

A MULTIPLE TESTING  
PROCEDURE  
TO ASSOCIATE MICROARRAY  
DATA  
WITH SURVIVAL

Sin-Ho Jung, Kouros Owzar, Stephen George

Department of Biostatistics and Bioinformatics

Duke University

## Approaches to identify genes which are prognostic for survival

- Dimension reduction using PCA plus Cox regression, Nguyen et al. (2001)
- Dichotomize by alive/dead, e.g. Shannon et al. (2002)
- Recurrence/no-recurrence within 2-year
- Marginal Cox model, and SAM e.g. Sorlie et al. (2001)
- Log-rank test for the most significant classification, and Bonferroni adjustment e.g. Jenssen et al. (2002)

## Notations

1.  $n$  subjects ( $i = 1, \dots, n$ )  
 $m$  genes ( $j = 1, \dots, m$ )
2. Gene expression data:  $(Z_{i1}, \dots, Z_{im})$
3. Survival data:  
For  $T_i$  = survival time,  $C_i$  = censoring time,  
 $X_i = \min(T_i, C_i)$ ,  $\Delta_i = I(T_i \leq C_i)$ .

## Procedure:

- Derive a measure for  $\text{assoc}(T_i, Z_{ij})$
- Test on  $H_j : \text{assoc}(T_i, Z_{ij}) = 0$
- Compute adjusted p-values

## Assoc( $T, Z$ ) - Jung et al. (1995)

Consider a single gene case.

1. Order  $Z_1, \dots, Z_n \Rightarrow Z_{(1)} < \dots < Z_{(n)}$
2. Suppose a large  $Z$  value  $\leftrightarrow$  high risk.
3. For  $k = 1, \dots, n - 1$ , using  $Z_{(k)}$  as a cutoff, partition  $n$  patients into
  - $\left\{ \begin{array}{l} \text{low-risk group: } Z_i \leq Z_{(k)} \quad (n_1 = k) \\ \text{high-risk group: } Z_i > Z_{(k)} \quad (n_2 = n - k) \end{array} \right.$
4. Compute the log-rank statistic  
 $U_k = \text{Low-High}$ .
5. A large positive  $U_k$  means  $Z_{(k)}$  is a good cutoff to partition between low- and high-risk groups.

6. As a general measure of association, use

$$\begin{aligned} W &= \sum_k U_k \\ &= \sum_{i=1}^n \int_0^\infty \left( R_i - \frac{\sum_{i'}^n R_{i'} Y_{i'}(t)}{Y(t)} \right) dN_i(t), \end{aligned}$$

$R_i$  is the rank of  $Z_i$  among  $Z_1, \dots, Z_n$ .

7. If there exist ties among  $Z_i$ 's, assign the average of the ranks that could be possibly taken by the tied observations.

## Property of $W$

1.  $W \approx 0$ , if  $T$  and  $Z$  are independent
2. A large positive  $W$  means large  $Z$  is associated with high risk.

A Single-Step Multiple Testing  
- Jung (2003)

$W_j$  for genes  $j = 1, \dots, m$ .

$H_j : T$  and  $Z_j$  are not associated.

$\bar{H}_j : T$  and  $Z_j$  are negatively associated.

Reject  $H_j$  if  $W_j > c_\alpha$ ,

$c_\alpha$  satisfying

$$\alpha = P\{\cup_{j=1, \dots, m} (W_j \geq c_\alpha) | H_0\}$$

$$= P(\max_{j=1, \dots, m} W_j \geq c_\alpha | H_0),$$

where  $H_0 = \cap_{j=1, \dots, m} H_j$ .

FWER =  $\alpha$

Need dist. of  $V = \max_{j=1, \dots, m} W_j$  under  $H_0$ .

## Permutation Method

Subj.	Survival	Gene
1	$(X_1, \Delta_1)$	$(Z_{11}, \dots, Z_{1m})$
$\vdots$	$\vdots$	$\vdots$
$i$	$(X_i, \Delta_i)$	$(Z_{i1}, \dots, Z_{im})$
$\vdots$	$\vdots$	$\vdots$
$n$	$(X_n, \Delta_n)$	$(Z_{n1}, \dots, Z_{nm})$

1. With the gene data fixed, permute the survival data.
2. Obtain  $(w_1^{(b)}, \dots, w_m^{(b)})$   
and  $v_b = \max_{j=1, \dots, m} w_j^{(b)}$   
from the  $b$ -th permutation ( $b = 1, \dots, B$ ).
3.  $c_\alpha \approx v_{[(1-\alpha)B]}$

4. (Unadjusted) p-value for gene  $j$  is

$$p_j \approx \frac{\sum_{b=1}^B I(w_j^{(b)} \geq w_j)}{B},$$

where  $w_j = \text{obsd } W_j$  from the orig. data.

5. Adjusted p-value for gene  $j$  is

$$p_j \approx \frac{\sum_{b=1}^B I(v_b \geq w_j)}{B}.$$

## Two-Sided Tests

$H_j$  :  $T$  and  $Z_j$  are not associated.

$\bar{H}_j$  :  $T$  and  $Z_j$  have some association.

Reject  $H_j$  if  $|W_j| > c_\alpha$ ,  
 $c_\alpha$  satisfying

$$\begin{aligned}\alpha &= P\{\cup_{j=1,\dots,m}(|W_j| \geq c_\alpha) | H_0\} \\ &= P(\max_{j=1,\dots,m} |W_j| \geq c_\alpha | H_0)\end{aligned}$$

Let  $V = \max_{j=1,\dots,m} |W_j|$ .

Adjusted p-value for gene  $j$  is

$$p_j \approx \frac{\sum_{b=1}^B I(v_b \geq |w_j|)}{B}.$$

## Simulations

1.  $m = 1,000$
2.  $T_i \sim LN(0, 1)$
3.  $(Z_{i1}, \dots, Z_{im}) \sim N\{0, CS(\gamma)\}$
4.  $\text{corr}(\log T_i, Z_{ij}) = \rho$  for  $1 \leq j \leq D$   
 $= 0$  for  $D + 1 \leq j \leq m$
5.  $C_i \sim U(f, a + f)$ ,  
 $a$  and  $f$  chosen for 20% or 40% censoring.
6. FWE,  $\alpha = .05$
7.  $B = 1000$  permutations  
 $N = 1000$  simulations.

Table 1. Empirical FWE ( $m = 1000, D = 0$ )

$n$	Censoring	$\gamma = 0$	.3	.6
20	20%	.055	.054	.052
	40%	.044	.035	.046
50	20%	.053	.048	.045
	40%	.053	.057	.054

Table 2. Empirical rejection rate of each  $H_j$   
(and global power.)

$n$	$\rho$	$D$	Genes	20% censoring		
				$\gamma = 0$	.3	.6
20	.3	5	$j \leq D$	.001-.006	.001-.003	.002-.009
			$j > D$	.000-.002 (.076)	.000-.002 (.062)	.000-.003 (.071)
		15	$j \leq D$	.000-.006	.000-.006	.003-.009
			$j > D$	.000-.002 (.084)	.000-.002 (.081)	.000-.003 (.092)
	.6	5	$j \leq D$	.042-.059	.051-.071	.098-.120
			$j > D$	.000-.001 (.259)	.000-.002 (.268)	.000-.003 (.307)
		15	$j \leq D$	.041-.066	.051-.072	.097-.129
			$j > D$	.000-.001 (.502)	.000-.002 (.448)	.000-.003 (.459)
50	.3	5	$j \leq D$	.010-.026	.019-.032	.048-.062
			$j > D$	.000-.001 (.119)	.000-.002 (.139)	.000-.004 (.176)
		15	$j \leq D$	.011-.029	.018-.034	.040-.068
			$j > D$	.000-.001 (.270)	.000-.002 (.234)	.000-.004 (.261)
	.6	5	$j \leq D$	.558-.578	.615-.631	.743-.753
			$j > D$	.000-.001 (.947)	.000-.002 (.921)	.000-.003 (.931)
		15	$j \leq D$	.536-.590	.604-.647	.729-.776
			$j > D$	.000-.001 (.999)	.000-.002 (.987)	.000-.003 (.976)

Example: Beer et al. (2002)

- Adeno Lung Cancer ( $n = 86$ )
- Stages I ( $n = 67$ ) and III ( $n = 19$ )
- Oligonucleotide arrays ( $m = 4966$ )
- Overall survival (24 deaths)

Some prognostic genes  
 ( $B = 10,000$ )

Gene	p-value	
	adj	unadj
Over-expressed in high-risk pts		
731	.7809	.0007
1167	.0426	.0001
1176	.5555	.0003
1517	.1976	.0000
2952	.6065	.0004
3800	.3148	.0001
4081	.3781	.0000
Under-expressed in high-risk pts		
382	.0125	.0000
772	.6767	.0003
1749	.3387	.0004
1858	.5421	.0000
1983	.2588	.0000
2573	.4986	.0001
3002	.7530	.0007
3926	.6863	.0004
3948	.6185	.0002