

ASSOCIATING MICROARRAY DATA WITH A SURVIVAL ENDPOINT

[Extended Abstract]

Sin-Ho Jung
Department of Biostatistics
and Bioinformatics
DUMC Box 3627
Durham, North Carolina 27710
sinho.jung@duke.edu

Kouros Owzar Department
of Biostatistics and
Bioinformatics
DUMC Box 3627
Durham, North Carolina 27710
kouros.owzar@duke.edu

Stephen George
Department of Biostatistics
and Bioinformatics
DUMC Box 3958
Durham, North Carolina 27710
stephen.george@duke.edu

ABSTRACT

In many microarray studies the primary objective is to identify, from large panel of genes, those which are prognostic markers of a censored survival endpoint such as time to disease recurrence or death. Often, these genes are considered prognostic in the sense that their respective expressions are associated, in an appropriate sense, with the survival endpoint of interest. From a practical point of view, this requires not only specifying a appropriate measure of association and a suitable statistic thereof, but also, as the number of genes is large, proper handling of the consequential issue of multiplicity. In this paper, we will address the aforementioned issues by utilizing a general correlation measure and a non-parametric statistic thereof, and by controlling the family-wise error rate by employing permutation resampling. Comprehensive simulation studies are conducted to investigate the statistical properties of the proposed procedure. The proposed procedure is demonstrated with microarray data.

Keywords

Censoring, Family-wise Error Rate, Rank Correlation, Multiple Testing

1. INTRODUCTION

In the early microarray studies, the primary objective has been focused, as for example in [5], on identifying genes which express differentially in different phenotypes. A new trend of microarray studies these days, however, is to discover the relationship between gene expression level and aggressiveness of a disease (such as cancer) or existence of tumor residue after tumor resection. The most popular and often useful endpoint in this type of studies may be time to a clinical event, such as disease recurrence or death. In

this context, a gene is often considered to be prognostic if its expression level is associated with the survival endpoint. The times to such events are usually subject to censoring due to loss to follow up or termination of the study.

When considering or devising a statistical method for analysis of such studies, the following issues need to be taken into account. First, one needs to choose a measure of association which properly quantifies the dependence, or for that matter the lack thereof, between the survival endpoint and each of a large number of genes. Secondly, one needs to specify a statistic which robustly estimates this measure of association for each gene. Finally, given that the number of genes under consideration is rather large, it is imperative to ensure that the overall error-rate is, in some appropriate sense, adequately controlled.

A heuristic approach to this end may be to partition the subjects into two groups: event versus no event, and proceed by using a standard approach, such as two-sample t-test statistic, to identify genes differentially expressing between the two groups (see for example [1] and [12]). This approach, however, can be biased as the subjects in the study usually have different follow up periods.

[11] and [9] reduce the dimension of gene expression data using a method like principal component analysis and fit a Cox's regression model using the derived components as covariates. This approach fails not only to test on the marginal correlation of a gene (or a principal component) and the survival variable but also to adjust for the multiplicity of the testing procedure.

[3] identified a prognostic gene with p-value calculated by fitting a Cox's regression model without adjusting for multiplicity of the original genes. [15] fit an univariate Cox regression model on each gene expression level and applied [4] approach to the resulting univariate (or unadjusted) p-values to adjust for the multiple testing procedure. [13] also fit univariate Cox's regression models on gene expression levels and applied a method called SAM ([14]) to discover prognostic genes.

For each gene, [6] sort the expression level observations and

partition all patients into two groups using each order statistic as a cutoff: one group for those patients who have gene expression levels smaller than the cutoff and the other for those who have gene expression levels equal to or larger than the cutoff. The (standardized) logrank statistic is calculated to compare the survival distribution between the two groups. They take the largest logrank statistic with respect to all possible cutoffs for each gene. Finally, they apply Bonferroni method to identify prognostic genes adjusting for multiple testing. They argue that the choice of maximum logrank test statistics causes anti-conservativeness, but it will be compromised by the conservative Bonferroni adjustment. This method does not provide an accurate control of the family-wise error rate (FWER).

In this paper, we will review a measure of rank correlation between a continuous variable and a survival variable. This measure was originally proposed by [10] and was subsequently used by [8] to compare two correlated surrogate markers which are prognostic for patient's survival time. We use this rank correlation measure to associate each gene expression level with a survival variable, and discover prognostic genes using a single-step multiple testing method outlined by [7], which uses a permutation method to derive adjusted p-values for the genes. Simulation studies are conducted to evaluate the performance of the proposed procedure. To demonstrate the applicability of the procedure to real microarray data a case study is presented.

2. MULTIPLE TESTING USING A RANK CORRELATION

At first, we investigate a rank correlation between the expression level of a gene (a continuous variable) and a survival endpoint. Suppose that there are n subjects. For patient i , T_i denotes the time to an event (such as tumor recurrence or death), called survival time hereafter. The survival time may be censored due to loss to follow-up or study completion, so that we observe $X_i = \min(T_i, C_i)$ together with censoring indicator $\Delta_i = I(T_i \leq C_i)$, where C_i is the censoring time which is assumed to be independent of T_i given gene expression level. Let $Y_i(t) = I(X_i \geq t)$ and $N_i(t) = \Delta_i I(X_i \leq t)$ be the at-risk and the death processes for patient i , respectively. Let $Y(t) = \sum_{i=1}^n Y_i(t)$.

Let m denote the number of genes under consideration and $(Z_{ij}, 1 \leq j \leq m)$ denote the expression levels of m genes from patient i . Usually the gene expression data within each subject are correlated.

As a general measure of association between the expression level for gene j and the survival data, we use

$$W_j = \sum_{i=1}^n \int_0^\infty (R_{ij} - \frac{\sum_{i'=1}^n R_{i'j} Y_{i'}(t)}{Y(t)}) dN_i(t), \quad (1)$$

where R_{ij} is the rank of Z_{ij} among (Z_{1j}, \dots, Z_{nj}) . Note that W_j has a form of covariance between R_{ij} and death process $N_i(t)$. W takes a large positive (negative) value if the gene tends to overexpress in the high (low) risk patients, and distribute around 0 if the gene expression does not have any impact on the survival.

W is rank-invariant with respect to Z as well as T . Fur-

thermore, W is the same as the score test based on Cox's partial likelihood for a proportional hazards model in which the rank of Z_i is used as a time-independent covariate, see [10].

[8] used this measure to compare two correlated markers ('genes' here) which are prognostic for survival time. Contrary to [10], we do not assume any (semi-)parametric model between survival and gene expression level in this paper.

We want to identify genes that are associated with survival time. We consider hypotheses,

$$H_j : T \text{ and } Z_j \text{ are not associated}, \quad (2)$$

versus

$$\bar{H}_j : T \text{ and } Z_j \text{ are negatively associated}, \quad (3)$$

i.e., gene j tends to overexpress in high-risk patients. Then, we may reject H_j in favor of \bar{H}_j for a large value of W_j . Let $H_0 = \cap_{j=1}^m H_j$, under which no genes are associated with survival time. Given FWER α , we want to find a common critical value c_α that satisfies

$$P\{\cup_{j=1, \dots, m} (Z_j \geq c_\alpha) | H_0\} = P(\max_{j=1, \dots, m} Z_j \geq c_\alpha | H_0) \leq \alpha. \quad (4)$$

In order to solve (4), we need to know the joint distribution of (Z_1, \dots, Z_m) under H_0 . However usually this is not available in a closed form due to the extremely high dimension of the random vector. So, we propose to use a permutation method to approximate the null distribution of the test statistics.

In order to maintain the correlation structure among m genes, we keep the m gene expressions (Z_{i1}, \dots, Z_{im}) together. We generate permutation data under H_0 by separating the survival data (X_i, Δ_i) from the gene expression data (Z_{i1}, \dots, Z_{im}) , and randomly matching the survival data with the gene expression data. For a permutation (j_1, \dots, j_n) of $(1, \dots, n)$, a permutation sample is generated as $\{(X_{j_i}, \Delta_{j_i}, Z_{i1}, \dots, Z_{im}), i = 1, \dots, n\}$. Since our test statistics depend on the gene expression data only through their rank, we may replace the gene expression data with their ranks, i.e. a permutation sample is given as

$$\{(X_{j_i}, \Delta_{j_i}, R_{i1}, \dots, R_{im}), i = 1, \dots, n\}.$$

From the b -th permutation sample, we calculate the test statistics $w_1^{(b)}, \dots, w_m^{(b)}$ and $\bar{w}^{(b)} = \max_{j=1}^m w_j^{(b)}$. The number of possible permutations, $n!$, as for example for a moderate sample size $n = 10$, we have $n! = 3,628,800$, is typically rather large. We may choose a reasonably large number of these permutations, say $B = 10,000$. Then, from (1), c_α is approximated by the $[B(1 - \alpha) + 1]$ -st order statistic of $\bar{w}^{(1)}, \dots, \bar{w}^{(B)}$, where $[a]$ is the largest integer that is smaller than a .

An adjusted p-value for gene j is defined as the minimum FWER at which H_j will be rejected. So, with an observed test statistic value $W_j = w_j$ for gene j , the adjusted p-value is given as

$$p_j = P(\max_{j'=1, \dots, m} W_{j'} \geq w_j | H_0), \quad (5)$$

which can be estimated from the permutations:

$$p_j \approx \frac{\sum_{b=1}^B I(\bar{w}^{(b)} \leq w_j)}{B}. \quad (6)$$

Jung (2003) investigated a similar testing procedure for multiple two-sample t-tests.

If we want to identify the genes either positively or negatively associated with survival time, then we may use two-sided tests. For marginal two-sided tests, we want to find a common critical value \tilde{c}_α that satisfies

$$P\left(\max_{j=1, \dots, m} |W_j| \geq \tilde{c}_\alpha | H_0\right) \leq \alpha. \quad (7)$$

We can approximate \tilde{c}_α using the same permutation method described above except that we obtain

$$\bar{w}^{(b)} = \max_{j=1, \dots, m} |W_j^{(b)}|$$

from the b -th permutation data. Adjusted p-value for gene j , with observed test statistic $W_j = w_j$, also should be modified as

$$p_j = P\left(\max_{j'=1, \dots, m} |W_{j'}| \geq |w_j| | H_0\right), \quad (8)$$

which is approximated as

$$p_j \approx \frac{\sum_{b=1}^B I(\bar{w}^{(b)} \leq |w_j|)}{B}. \quad (9)$$

Given FWER α , we may reject H_j if $W_j > c_\alpha$ or $p_j < \alpha$. Calculation of c_α involves sorting of $(\bar{w}^{(b)}, 1 \leq b \leq B)$, so that the testing procedure using adjusted p-values requires less computing time.

3. NUMERICAL STUDIES

We investigate the performance of our new multiple testing procedure with a larger number of genes, m . We generate gene expression data from a multivariate normal distribution and survival time from a lognormal distribution, which is negatively correlated with prognostic genes. In type I error analyses, we generate the data as follows. For iid $N(0,1)$ random numbers $\tau_i, \epsilon_{i0}, \epsilon_{i1}, \dots, \epsilon_{im}$, we set

$$\begin{aligned} \log(T_i) &= \tau_i \\ Z_{ij} &= \epsilon_{ij} \sqrt{1-\rho} + \epsilon_{i0} \sqrt{\rho} \quad \text{for } 1 \leq j \leq m. \end{aligned}$$

Then, the survival time is not associated with any genes, and the gene expression data have a multivariate normal distribution with zero means, unit variances and a compound symmetric correlation structure with coefficient ρ . We consider $m = 1,000$, $n = 20$ or 50 , $\rho = 0, .3$ or $.6$, and 20% or 40% censoring. A censoring time is generated from $U(0, c_0)$ with c_0 chosen for 40% censoring. With c_0 fixed at this value, a censoring variable for 20% censoring is generated from $U(c_1, c_0 + c_1)$ by choosing a proper c_1 value. Null distribution of the test statistic is approximated from $B = 1,000$ random sample of $n!$ possible permutations. Empirical FWER is computed as the proportion of samples rejecting H_0 by our testing procedure with one-sided FWER=.05 among $N = 1,000$ simulations. Simulation results are reported in Table 1. Our procedure overall has an empirical FWER close to the nominal level.

For power analyses, the first D genes are set to be prognostic with correlation coefficients r with $\log(T)$. The data are generated as follows. For iid $N(0,1)$ random numbers $\tau_{i0}, \tau_i, \epsilon_{i0}, \epsilon_{i1}, \dots, \epsilon_{im}$, we obtain

$$\log(T_i) = \tau_i \sqrt{1-r} - \tau_{i0} \sqrt{r}$$

$$Z_{ij} = \begin{cases} \epsilon_{ij} \sqrt{1-\rho} + \epsilon_{i0} \sqrt{\rho} + \tau_{i0} \sqrt{r} & \text{for } 1 \leq j \leq D \\ \epsilon_{ij} \sqrt{1-\rho} + \epsilon_{i0} \sqrt{\rho} & \text{for } D+1 \leq j \leq m \end{cases}$$

It can be shown that $\text{corr}(\log T_i, Z_{ij}) = -r/\sqrt{1+r} \equiv \eta$ for $1 \leq j \leq D$ and $= 0$ for $D+1 \leq j \leq m$; $\text{corr}(Z_{ij}, Z_{ij'}) = (\rho+r)/(1+r)$ for $1 \leq j < j' \leq D$, $= \rho/\sqrt{1+r}$ for $1 \leq j \leq D < j' \leq m$ and $= \rho$ for $D+1 \leq j < j' \leq m$. Note that η is the parameter of interest. We set $n = 50$, $D = 5, 10$ or 15 ; $\eta = .3$ or $.6$ in addition to the parameters set for the type I error analyses. The simulation results are summarized in Table 2. For non-prognostic genes the false discovery rates, i.e. the probability that H_j is rejected when H_j is true, are very low. Overall power, i.e. the probability that any H_j is rejected, and true discovery rate, i.e. the probability that H_j is rejected when H_j is true, increase in η . With $\eta = .3$, overall power and true discovery rate are low. But with $\eta = .6$, power and true discovery are very high. True discovery rate increases in ρ , but power does not seem to change in ρ .

[2] used oligonucleotide arrays to generate gene expression data for $m = 4966$ genes from $n = 86$ patients with lung adenocarcinoma. We applied our multiple testing method to their data to identify prognostic genes. Analysis results are summarized in Table 3. The columns with $\rho < 0$ ($\rho > 0$) are for testing the one-sided alternative hypotheses that a gene tends to overexpress in high (low) risk patients. The columns $\rho \neq 0$ are for two-sided tests. Adjusted and unadjusted p-values are listed for those genes with either one-sided adjusted p-value smaller than 0.8. We observe that Gene 385 (BCL2) underexpresses and Gene 1167 (BIRC5) overexpresses in high risk patients. For all other genes listed in table 6, the unadjusted p-values are very small. The corresponding adjusted p-values for these genes, however, are not small enough for statistical significance after adjusting for multiplicity of the testing procedure.

4. CONCLUSIONS

What has been presented and discussed in this paper, is a comprehensive non-parametric procedure for analyzing microarray studies whose primary endpoint is a censored survival variable. For a method to be useful in microarray data analysis, it must address the following three issues:

- i. the ability to quantify the degree of association and the corresponding statistical significance between *each* gene and the survival variable;
- ii. the ability to control the *overall* error rate;
- iii. robustness against outliers and model misspecification.

As illustrated in the literature review presented in the introductory section, there is a sizable literature on analyzing microarray studies whose primary endpoint is a censored survival variable. What sets the method proposed in this paper

apart, is the fact that it *simultaneously* addresses all of the three aforementioned issues. Furthermore, as this method is inferential, rather than data-driven, it will not only be useful from the point of view of exploratory data analysis, but should also serve as an invaluable tool for sample size and power calculations in designing experiments for which microarray studies with survival endpoints are planned.

To demonstrate the performance as well as applicability of the method, we have presented simulation as well as case studies in section 4. The simulation study suggest that false-rejection rate (i.e., incorrectly declaring a non-prognostic gene as prognostic) for this method is virtually negligible. For moderately sized studies (e.g., $n = 50$ in this case), the method will have very good global power (i.e., probability of detecting at least one of the prognostic genes) as long as the hypothesized effect size is reasonably large (e.g., $\eta = 0.6$ in this case). In those cases, the method also enjoys good true-discovery rates (i.e., correctly declaring a prognostic gene as prognostic). Furthermore, the results in table 1 illustrate the ability of the method to adequately control the FWER.

The amount of association between the survival endpoint and the expression level of gene j was quantified estimated by W_j . One can generate variations of the proposed method by employing other types of association measures and statistics. Such extensions are subject to active pursuit by the authors.

5. REFERENCES

- [1] A. André, T. Karn, C. Solbach, T. Seiter, K. Strebhardt, U. Holtrich, and M. Kaufmann. Identification of high risk breast-cancer patients by gene expression profiling. *Lancet*, 359:131–132, 2002.
- [2] D. G. Beer, S. L. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 8:816–824, 2002.
- [3] S. M. Dhanasekaran, T. R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan. Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412(6849):822–6, 2001.
- [4] S. D. Dubey. Adjustment of p -values for multiplicities of intercorrelating symptoms. In *Statistics in the Pharmaceutical Industry (Second Edition)*, pages 513–527, 1993.
- [5] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and L. E. S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(15):531–537, 1999.
- [6] T. K. Jenssen, W. P. Kuo, T. Stokke, and E. Hovig. Associations between gene expressions in breast cancer and patient survival. *Hum Genet*, 111:411–20, 2002.
- [7] S. H. Jung. Single step multiple testing. submitted, 2003.
- [8] S. H. Jung, S. Wieand, and S. S. Cha. A statistic for comparing two correlated markers which are prognostic for time to an event. *Statistics in Medicine*, 14:2217–2225, 1995.
- [9] D. V. Nguyen and D. M. Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50, 2002.
- [10] J. O’Quigley and R. L. Prentice. Nonparametric tests of association between survival time and continuously measured covariates: The logit-rank and associated procedures. *Biometrics*, 47:117–127, 1991.
- [11] P. J. Park, T. L., and K. I. S. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, 18:S120–S127, 2002.
- [12] W. D. Shannon, M. A. Watson, A. Perry, and K. Rich. ntel statistics to correlate gene expression levels from microarrays with clinical covariates. *Genetic Epidemiology*, 23:87–96, 2002.
- [13] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, J. S. S., T. Thorsen, H. Quist, M. J. C. B. P. O., D. Botstein, P. Eystein Lonning, and B.-D. A. L. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS*, 98(19):10869–74, 2001.
- [14] V. G. Tusher and R. Tibshirani. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98(9):5116–21, 2001.
- [15] D. A. Wigle, I. Jurisica, N. Radulovich, M. Pintilie, J. Rossant, N. Liu, C. Lu, J. Woodgett, I. Seiden, M. Johnston, S. Keshavjee, G. Darling, T. Winton, B.-J. Breitkreutz, P. Jorgenson, M. Tyers, F. A. Shepherd, and M. S. Tsao. Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res*, 62(11):3005–3008, 2002.

6. TABLES

Table 1: Empirical FWER for nominal 5% FWER with $m = 1,000$, $B = 1,000$ and $N = 1,000$.

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

Table 2: Empirical rejection rate of each H_j under $n = 50$, $m = 1,000$, $B = 1,000$ and $N = 1,000$. Genes are grouped for prognostic ones ($j = 1, \dots, D$) and non-prognostic ones ($j = D + 1, \dots, m$). The numbers in parentheses are empirical rejection rate of any of these hypotheses, called global power.

η	D	Censoring	Genes	$\rho = 0$.3	.6	
.3	5	20%	$j \leq D$.001-.006	.001-.003	.002-.009	
			$j > D$.000-.002 (.076)	.000-.002 (.062)	.000-.003 (.071)	
		40%	$j \leq D$.000-.003	.001-.004	.003-.012	
			$j > D$.000-.002 (.049)	.000-.002 (.063)	.000-.004 (.079)	
		15	20%	$j \leq D$.000-.006	.000-.006	.003-.009
				$j > D$.000-.002 (.084)	.000-.002 (.081)	.000-.003 (.092)
	40%		$j \leq D$.000-.003	.001-.007	.003-.012	
			$j > D$.000-.002 (.061)	.000-.003 (.077)	.000-.004 (.090)	
	.6	5	20%	$j \leq D$.042-.059	.051-.071	.098-.120
				$j > D$.000-.001 (.259)	.000-.002 (.268)	.000-.003 (.307)
			40%	$j \leq D$.024-.043	.035-.048	.069-.090
		$j > D$.000-.002 (.186)	.000-.002 (.199)	.000-.003 (.228)	
15		20%	$j \leq D$.041-.066	.051-.072	.097-.129	
			$j > D$.000-.001 (.502)	.000-.002 (.448)	.000-.003 (.459)	
	40%	$j \leq D$.023-.044	.030-.050	.070-.096		
$j > D$.000-.002 (.343)	.000-.002 (.322)	.000-.004 (.347)			

Table 3: Analysis results for Michigan Data ($n = 86$, $m = 4966$) with $B = 10,000$ permutations. Genes with at least one adjusted one-sided p-value smaller than .8 are listed. (– meaning a one-sided adjusted p-value of 1.0000)

Gene	Adjusted p-value			Unadjusted p-value		
	$\rho < 0$	$\rho > 0$	$\rho \neq 0$	$\rho < 0$	$\rho > 0$	$\rho \neq 0$
382	–	.0125	.0227	–	.0000	.0000
485	.7131	–	.8794	.0006	–	.0009
670	–	.7714	.9145	–	.0004	.0011
731	.7809	–	.9218	.0007	–	.0014
772	–	.6767	.8490	–	.0003	.0006
1167	.0426	–	.0769	.0001	–	.0001
1176	.5555	–	.7504	.0003	–	.0005
1517	.1976	–	.3229	.0000	–	.0002
1604	–	.7423	.8961	–	.0007	.0010
1749	–	.3387	.5101	–	.0004	.0004
1858	–	.5421	.7314	–	.0000	.0003
1875	.6509	–	.8300	.0002	–	.0005
1916	–	.7022	.8687	–	.0007	.0010
1983	–	.2588	.4099	–	.0000	.0002
2573	–	.4986	.6919	–	.0001	.0003
2623	.3894	–	.5720	.0001	–	.0002
2952	.6065	–	.7926	.0004	–	.0010
3002	–	.7530	.9043	–	.0007	.0009
3249	–	.4731	.6666	–	.0002	.0006
3328	–	.7880	.9236	–	.0009	.0014
3503	–	.7936	.9254	–	.0009	.0014
3800	.3148	–	.4847	.0001	–	.0004
3926	–	.6863	.8571	–	.0004	.0008
3948	–	.6185	.8007	–	.0002	.0003
4081	.3781	–	.5570	.0000	–	.0000
4665	–	.2987	.4593	–	.0000	.0000