

Entropy and Survival-based Weights to Combine Affymetrix Array Types in the Analysis of Differential Expression and Survival

^{1*}Jianhua Hu, ²Guosheng Yin, ²Jeffrey S. Morris, ²Li Zhang, ¹Fred A. Wright

¹Department of Biostatistics
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599

²Department of Biostatistics
M. D. Anderson Cancer Center, University of Texas
Houston, TX 77030

*Tel: 919-969-6680, Email: jhu@bios.unc.edu

ABSTRACT

In order to comprehensively identify genes with expression levels that correlate with survival for patients with lung adenocarcinoma, we combined data across the Harvard and Michigan studies. Two different versions of Affymetrix oligonucleotide microarrays were used in these two studies. We propose combining arrays of different platforms by assigning weights to the expression levels of each gene across data sets based on the entropy of the residual matrix. In each data set, the expression level of each gene is quantified by the “reduced” model proposed by Li and Wong [9], which is equivalent to a method using the singular value decomposition [7]. We combined information across different chip types by first identifying common genes on the two chip types, and then assigning weights based on residual entropy for each gene. To incorporate clinical information, especially survival data, in detecting important genes, we propose a new method based on weighted t-tests (WTT). The survival information can be absorbed into a set of weights assigned to the expression intensities across all the arrays or subjects, based on the predicted median survival time using the Cox proportional hazards model [5]. Important genes can be identified by comparing the survival-weighted t-tests with another t-test comparing the cancer patients to the reference group, and error rates can be controlled by permutation procedures.

Keywords: Cox model, Entropy, False discovery rate, Median survival time, Singular value decomposition, Weighted t-test.

1. INTRODUCTION

DNA microarray technology has been increasingly used and now plays an important role in many areas of biomedical research. This technology allows us to monitor the expression levels of very large numbers of genes simultaneously and repeatedly in cell lines, human tissues and a wide range of organisms.

Multiprobe oligonucleotide arrays provide useful redundancy in interrogating the same transcript via multiple probes, which are complementary to different regions of the mRNA. The perfect match (PM) probes are exactly complementary to the prepared

target cRNA, while mismatch (MM) probes identical to the corresponding PM probe, except for a base change at the 13th position. The comparison of PM vs. MM signal intensity is designed to give information about nonspecific cross-hybridization or other measurement errors.

The two oligonucleotide array studies were chosen for exploration. The Michigan data set [2] uses the Hu6800 platform (20 probe pairs, 7,129 probe sets.) The Harvard [4] data set uses the newer type U95Av2 platform (16 probe pairs, 12,625 probe sets). A common objective of the two studies was to identify important genes with expression related to lung adenocarcinoma, a leading cause of cancer deaths in the United States, and genes whose expression was related to patient survival.

Our research objective was to further combine information from the two different studies, identifying genes that were differentially expressed in non-cancer samples vs. histologically-defined lung adenocarcinoma samples. The latter was the most common histology was accompanied by relatively complete survival data. We propose a novel method to combine the two gene expression data sets, based on the singular value decomposition (SVD) expression estimates and residual entropy. Moreover, we propose a new weighted t-test for incorporating the clinical information into the procedure of identifying important genes.

2. EXAMINING SURVIVAL AND GENE EXPRESSION DATA

2.1 Survival Data

To examine the homogeneity of the two populations across the Harvard and Michigan studies, we started by comparing the clinical variables (survival data). Patient data from the two studies had comparable distributions of age, sex, and smoking status. However, only tumors of stages 1 and 3 were represented in the Michigan study, while tumors of stages 1, 2, 3 and 4 were represented in the Harvard data. We dichotomized the stage variable by combining the local stages (1 and 2) and the advanced stages (3 and 4). Figure 1 contains the Kaplan-Meier curves for the two studies. There is a significant difference in survival

between the two studies based on the log-rank test (p-value = 0.01). We included an indicator variable to account for an institution effect in the analysis, and otherwise the populations seemed comparable for a common pooled analysis.

2.2 Gene Expression Data

We log-transformed raw intensities of gene expression for each array and plotted them to remove bad chips. Samples L54, L88, L89, and L90 in the Michigan arrays contained a large round dark spot at the center of the chip, and samples L22, L30, L99, L81, L100, and L102 contained a large number of extremely bright outliers according to MAS5.0 (Affymetrix, Inc.). Two outlier chips were detected and removed in the Harvard dataset using dChip (CL2001040304 and CL2001041716). We kept the most recently dated run among the Harvard samples with 48 replicate arrays (the arrays had been duplicated due to a bad first run).

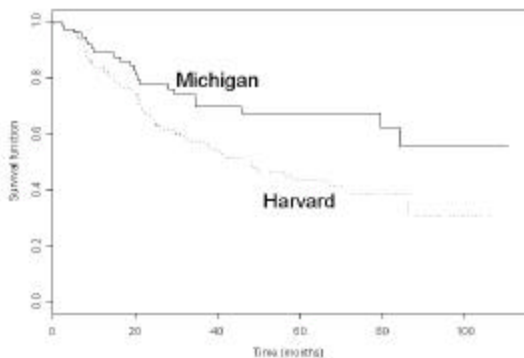


Figure 1: Estimated Kaplan-Meier survival curves.

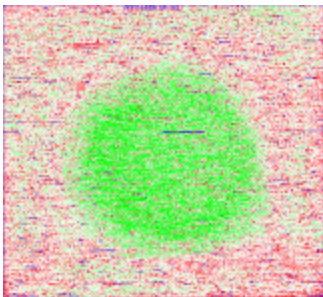


Figure 2: Image plot of log-expression for Sample L88 in Michigan Data Set. Green and red indicate log-expression levels below and above the median for the chip, indicating a bad chip. Samples L54, L88, L89, and L90 have similar plots.

This preprocessing resulted in a data set with matching clinical and microarray data for 229 patients, which includes control and primary lung adenocarcinomas samples, 143 from Harvard with 17 references, and 86 from Michigan with 10 references.

3. NORMALIZATION AND EXPRESSION INDEX ESTIMATION

Microarray normalization is important to remove sources of systematic variation in microarray expression estimates. Because scanned images may have a different level of overall brightness, it is important to normalize arrays such that they have comparable levels of brightness before analyzing gene expression levels.

Although a number of normalization techniques have been proposed, after exploratory analysis we chose simple linear normalization, using a synthetic “median array” as a reference. The median array consists of the median gene expression measurements for each gene across the arrays.

The term “expression index” describes a statistic used to represent an expression level for a gene. In recent years, several statistical methods for expression index computation for Affymetrix arrays proposed. These have included nonparametric approaches and parametric models. A multiplicative model [9] is feasible and popular, and has been shown to be superior to standard commercial software [8]. We performed the Li-Wong reduced model (LWR) using the SVD technique [7], because the latter lends it self to the entropy computation described below. Using the difference data matrix for each gene $PM_{i \times J} - MM_{i \times J}$, the first characteristic mode [6] of the singular value decomposition is proportional to the corresponding LWR estimates [7]. Here I and J denote the numbers of arrays and probes, respectively. The new method is more efficiently and closely related to the method of combining different platforms that is described below.

4. COMBINING DATA FROM DIFFERENT AFFYMETRIX ARRAYS

Determining how to combine the different types of Affymetrix oligonucleotide chips in the two studies was a main challenge. A list of common probe sets representing the same gene between these two different chip types in the Harvard and Michigan studies is available at the following dChip URL:

http://www.biostat.harvard.edu/complab/dchip/info_file.htm#common_probeset_file.

There are 5,987 probe set pairs representing the same genes across the two studies. However, due to differences in probe densities and probe sequences, the expression levels of the genes in these two chip types are not directly comparable. In order to obtain comparable gene expression levels across the two chip types, we introduced a technique for assigning weights to each expression index in the two data sets.

An important concept involved in our approach is entropy [10]. The entropy $H(f)$ of an absolutely continuous density $f(x)$ is defined $H(f) = \int f(x) \log f(x) dx$. H can be viewed as a measure of randomness or unpredictability of a random variable X , and has been applied in a variety of hypothesis testing problems. Some papers, e.g. Vasicek [12], discussed entropy-based hypothesis tests of normality or uniformity. Another important application of entropy is in combination with SVD for genome-

wide expression data [1]. Using ideas from [1], we define “fraction of eigenintensity” as

$$p_j = \frac{\mathbf{s}_j^2}{\sum_{j=1}^J \mathbf{s}_j^2}$$

where J is the number of probes and \mathbf{s}_j denotes the i th eigenvalue from the SVD decomposition. It indicates the degree of structure in the data matrix that can be captured by the i th eigenvector for arrays and probes. The discrete analogue of the Shannon entropy of a given data set is

$$e = \frac{-1}{\log(J)} \sum_{j=1}^J p_j \log(p_j)$$

where the entropy is scaled so that $0 \leq e \leq 1$. e describes the “randomness” of the data matrix, in the sense that SVD cannot meaningfully discern structure in fitting the data. In particular, $e=0$ corresponds to an ordered and redundant data set where all the expression is captured by a single eigenvalue, and $e=1$ corresponds to a disordered and random data set.

Assuming that the LWR is the true model from which the underlying expression index can be estimated, there should be no systematic pattern left in the residual matrix after fitting the model. This is equivalent to subtracting the product of the first set of eigenvalues of the data matrix and two eigenvectors from it using the SVD. The randomness of the residual matrix can be assessed by the distribution of its eigenvalues, quantified by the entropy. We reasoned that the data that better fits the model should have a higher entropy. First, in each study, the expression intensity matrix of each gene was standardized to a mean of 0 and a variance of 1 (to avoid one source of bias in the SVD). After applying the SVD, we obtained the eigenvalue entropies of the residual matrices for each gene. The distribution of the entropies across all the common genes in each data set is shown in Figure 3. Overall, the Harvard data appears much better, with entropies centered around 0.9, while those from Michigan are widely spread from 0 to 1. However, a few genes from the Harvard study were assigned very low weights (some even close to 0).

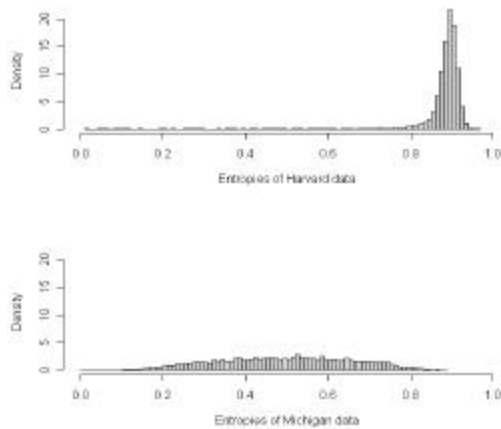


Figure 3: Distributions of entropies in the Harvard and Michigan studies.

The two studies have different dimensions in their data matrices for each gene. However, this fact has little impact on the entropies of the residual matrix, as demonstrated in some limited simulations. For each gene, the two entropy values (Harvard and Michigan) were then standardized to reach a sum of 1, and then within each study the appropriate weight was multiplied by the expression index to obtain a new entropy-weighted expression index. The weight is proportional to the entropy value, with a larger weight being assigned to the model-based expression index estimate in the study that has higher entropy for the specific gene.

To assess the performance of the entropy weighting strategy, we used the false discovery rate (FDR) as a comparison criterion. The FDR is defined as the expected proportion of false rejections (truly null) among the rejected hypotheses [3]. We followed the permutation procedures as implemented in the software SAM [11] to estimate the FDR. We computed ordinary t statistics based on both the unweighted and weighted expression data, and also conducted 1000 permutations of the t-tests. In each permutation, we randomly drew 27 samples from the total of 229 patients to be treated as the “reference” and treated the rest as the “cancer” patients. We estimated the FDR as a function of the number of genes that can be detected. Figure 4 shows the relationship between the FDR and the number of rejected genes up to 300. Clearly, the weighted data yielded a dramatically lower FDR level than the unweighted one.

Moreover, we examined the correlation among the samples. Ideally, the correlations within the reference or disease samples should be higher than between the reference and disease samples, if indeed gene expression can be used to

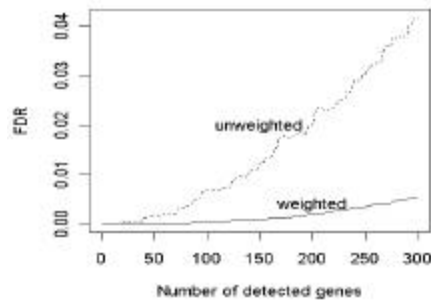


Figure 4: Comparison of FDRs between weighted and unweighted expression data.

discriminate between the groups. Examining the within-reference and within-disease samples in each data, we found the weighted method can increase the correlations over the unweighted. We calculated the differences of the pairwise correlations between the weighted and unweighted expression data, where 26.5% differences are between 0.1 and 0.5. The rest of the correlation differences vary around 0. We also compared the correlations between the reference and disease groups across the two data sets, and found that 78.8% pairwise correlations are lower in the

weighted expression data, though the differences were not dramatic.

5. IDENTIFYING IMPORTANT GENES

5.1 Weighting Based on Survival Data

Another major goal in this analysis is to combine the gene expression data with the patient survival data. To find those genes that either directly affect or can help predict the patients' survival, we need to take into account other clinical information such as survival time, censoring indicators, tumor stage, age, sex, and smoking status. Clearly, the Cox proportional hazards model is readily applicable. However, adding both the clinical variables and the gene data into the Cox model may cause a high-dimensionality problem. To address this issue, we introduce a new method of the weighted t-test (WTT). To incorporate the clinical information, some form of weight needs to be constructed for gene expression intensity data. Due to censoring, the construction of appropriate weights for each subject is quite challenging. In order to obtain reasonable weights, we propose using the predicted median survival time, as described below.

A total of 229 subjects are in the pooled sample, 188 are cancer patients with available recorded survival information. Our analysis includes the institution, age, sex, smoking status, and tumor stage as covariates. The institution is examined because the survival curves are very different between the studies performed at Harvard and at Michigan. For the i th subject with a covariate vector Z_i , the Cox proportional hazards model is given by

$$I(t | Z_i) = I_0(t) \exp(\mathbf{b}^T Z_i),$$

where $I_0(t)$ is the unknown and unspecified baseline hazard function and \mathbf{b} is the regression parameter of interest. For the i th subject, the survival function is given by

$$S(t | Z_i) = \exp\{-\Lambda_0(t) \exp(\mathbf{b}^T Z_i)\},$$

where $\Lambda_0(t)$ is the cumulative baseline hazard function.

Table 1: Parameter estimates under the Cox proportional hazards model (H.R. is the hazard ratio and S.E. is the standard error).

Covariate	$\hat{\mathbf{b}}$	H.R.	S.E.	p-value
Institution	0.6392	1.89	0.2501	0.011
Age	0.0267	1.03	0.0120	0.027
Sex	0.1292	1.14	0.2288	0.570
Smoking Status	0.0063	1.01	0.0032	0.048
Tumor Stage	1.5552	4.74	0.2666	<0.001

We constructed the predicted survival curve for each subject based on the clinical information only, from which we estimated the median survival time.

$$m_i = \inf\{t : S(t | Z_i) < 0.5\}.$$

We assigned an averaged median survival time to those subjects with missing survival information. For a given subject, no matter if an observation is a failure or is censored, m_i is determined by the covariate Z_i , which circumvents the potential bias caused by the censored data. Based on the predicted median survival time, we

calculated the weights that are proportional to m_i , for each cancer patient accordingly,

$$w_i = \frac{m_i}{\sum_{i=1}^n m_i} \times n$$

Moreover, for subjects in the reference sample, we did not assign any weights because they were controls and were all alive at the end of the study. Thus, we computed the weights using all the common clinical variables provided in the two studies.

With the survival-weighted expression data, we conducted a two-sample t-test for each gene to measure the difference in expression levels between the control group and the cancer patients. We also carried out t-tests on the expression data with no weight adjustment. In order to find the genes related to survival information, we examined the difference between the t-test statistics after and before the survival-weight adjustment, i.e., $d_k = t_{\text{after}} - t_{\text{before}}$, for the k th gene, $k=1, \dots, 5,987$. Note that by subtracting the t_{before} values, d_k was constructed to be sensitive to effects of expression on survival, and not on mere differences in expression in cancer vs. reference. We again performed 1,000 permutations. In each permuted dataset, we implemented the ordinary t-test and survival WTT, and recorded their statistics together with d_k for each gene. We ordered d_k for each permutation, and let $d_{(k)}$ denote the ordered d_k . Then, we calculated the averaged order statistics, $d_{(k)}$, across all the 1000 permutations. A gene is claimed to be related to survival when $d_{(k)} - d_{(k)}$ (if $d_{(k)}$ is positive) is larger than an appropriate threshold, or when $d_{(k)} - d_{(k)}$ (if $d_{(k)}$ is negative) is smaller than some threshold. We thus obtained a list of the most significant genes.

To accommodate the multiple testing issue in our analysis, again we applied the FDR criterion and identified the 12 gene most significantly related to survival as described above, while controlling for the FDR at 0.05. Furthermore, the statistical significance of the detected genes can be measured by p-values obtained from the permutation procedure, defined as the proportion of the difference d_k at least as extreme as that observed. A list of the 12 genes is shown in Table 2, along with the names of probe sets in U95a and Hu6800 platforms, gene descriptions and corresponding p-values. Here, as with Table 3 described below, we found an intriguing number of sex-specific genes and hormones. Sex was specifically included in the Cox model, and these results suggest taking a closer look at expression of these genes within each sex, and also for any possible lack of proportional hazards across the two sexes. Several other genes, including ribosomal proteins and those involved in immune response and cell differentiation, are typical of broad functional characteristics that have appeared in other cancer studies.

5.2 Differentiating between reference and cancer subjects

The WTT method can be used to identify the important genes that are differentially expressed in the two groups of reference and cancer patients. We have more confidence in choosing for further biological validation the genes found to have significant results

under both tests. The rationale is that such genes show both a difference between cancer vs. reference and also have an apparent effect on survival. Again, we used the SAM-like procedure to identify significant positive genes. By sorting $|d_{(k)}-d_{(k)}|$ and taking into consideration the sign of $d_{(k)}$ in both the ordinary t-test and WTT, the 15 most significant genes with the smallest sums of ranks of $|d_{(k)}-d_{(k)}|$ across the two t-test statistics are identified. Table 3 shows the names of the 15 probe sets in U95a and Hu6800 platforms, gene annotations and the ranks using the two different statistics.

6. CONCLUSIONS

In this study, we conducted the expression data analysis by using the SVD-based analogue of the LWR model. We imposed a SVD entropy weight on the expression of each gene, thereby demonstrably lowering the FDR in comparison of cancer vs. reference samples. The approach of using residual entropy to judge the quality of expression estimates can be applied in a much more general context. We incorporated survival data by imposing another weighting scheme based on the predicted median survival time to each subject. To identify important genes having significant impact on patient survival, we compared a survival weighted t-test to the corresponding ordinary t-test, with both tests using the entropy-weighted combined expression values.

Table 2: List of 12 most important genes related to survival.

U95A	Hu6800	Gene Annotation	p-value
725_i_at	J03071_cds3_f	Chorionic Somatomammotropin Hormone Cs-5	<0.001
35281_at	U31201_cds1	laminin, gamma 2 (nicein, kalinin, BM600, Herlitz junctional epidermolysis bullosa)	<0.001
34961_at	M88282	T cell activation, increased late expression	<0.001
31838_at	U79274	protein predicted by clone 23733	<0.001
37174_at	D14660	mitochondrial ribosomal protein L19	0.035
530_at	U16258	ribosomal protein S7	0.005
41643_at	X83301_s	Cluster Incl. X83301:H.sapiens SMA5 mRNA /cds=(319,741) /gb=X83301 /gi=603029	<0.001
35894_at	X14362	complement component (3b/4b) receptor 1, including Knops blood group system	<0.001
32864_at	L10102_rna1	sex determining region Y	0.002
32686_at	D86096_cds6	prostaglandin E receptor 3 (subtype EP3)	<0.001
722_at	D87957	rcd1 (required for cell differentiation, S.pombe) homolog 1	0.009
1338_s_at	X13930_f	X13930 /FEATURE=cds Human CYP2A4 mRNA for P-450 IIA4 protein	0.008

Ordered from the most significant to the least using SAM.

Table 3: The 15 most significant genes differentiating between reference and cancer groups.

U95A	Hu6800	Gene Annotation	rank($ d_{(k)}-d_{(k)} $) in t-test	rank($ d_{(k)}-d_{(k)} $) in WTT
725_i_at	HG1751-HT1768	Chorionic Somatomammotropin Hormone Cs-5	1	1
33780_at	M36200	vesicle-associated membrane protein 1 (synaptobrevin 1)	5	2
40081_at	HG3945-HT4215	phospholipid transfer protein	3	8
220_r_at	S76756_s	S76756 4R-MAP2=microtubule-associated protein, isoform	12	4
35281_at	U31201_cds1	laminin, gamma 2	2	15
38150_at	U22233	methylthioadenosine phosphorylase	8	10
32461_f_at	HG3137-HT3313	zinc finger protein 81 (HFZ20)	13	6
37263_at	U55206	gamma-glutamyl hydrolase (conjugase, folylpolygammag-l-h)	11	13
37975_at	X04011	cytochrome b-245, beta polypeptide (granulomatous disease)	17	9
37399_at	D17793	aldo-keto reductase family 1, member C3 (3-alpha h-d)	21	7
36287_at	X83368	phosphoinositide-3-kinase, catalytic, gamma polypeptide	18	11
1197_at	D00654	D00654 / DEFINITION=HUMACTSG7 Homo sapiens gene	25	5
1482_g_at	L23808	matrix metalloproteinase 12 (macrophage elastase)	27	3
36617_at	HG3342-HT3519_s	inhibitor of DNA binding 1, dominant negative h-l-h protein	20	12
35462_at	U17033	phospholipase A2 receptor 1, 180kD	10	32

Ordered from the most significant to the least according to the sum of ranks using SAM.

Moreover, we identified genes that were differentially expressed between the reference and cancer groups. Clearly, the proposed method can be extended to more general situations, such as an F-test in an ANOVA of multiple groups.

7. ACKNOWLEDGEMENTS

We thank Kevin Coombes and Fei Zou for helpful discussions.

8. REFERENCES

- [1] Alter, O., Brown, P. O. and Botstein, D. Singular Value Decomposition for Genome-wide Expression Data Processing and Modeling. *PNAS*, 97, 10101-10106, 2000
- [2] Beer, D, et al. Gene-expression Profiles Predict Survival of Patients with Lung Adenocarcinoma. *Nature Medicine*, 9 (816), 816-824, 2002.
- [3] Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B*, 57, 289-300, 1995.
- [4] Bhattacharjee, A, et al. Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses. *PNAS*, 98 (24), 13790-13795, 2001.
- [5] Cox, D. R. Regression Models and Life Tables (with discussion). *Journal of Royal Statistical Society, Series B*, 34, 187-220, 1972.
- [6] Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R. and Fedoroff, N. V. Fundamental Patterns Underlying Gene Expression Profiles: Simplicity from Complexity. *PNAS*, 97, 8409-8414, 2000.
- [7] Hu, J., Wright, F. A. and Zou, F. An Adaptive SVD Approach of Estimating Expression Indexes for Oligonucleotide Arrays. *Manuscript*, 2003.
- [8] Lemon, W. J., Palatini, J. J., Krahe, R. and Wright, F. A. Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics*, 18, 1470-1476, 2002
- [9] Li, C. and Wong, W. H. Model-based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection. *PNAS*, 98, 31-36, 2001a.
- [10] Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* 27, 379-423, 623-656, 1948. Reprinted in *Key Papers in the Development of Information Theory* (1974), ed. D. Slepian, New York: IEEE press, 5-2.
- [11] Tusher, V. G., Tibshirani, R. and Chu, G. Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. *PNAS*, 98, 5116-5121, 2001.
- [12] Vasicek, O. A Test for Normality Based on Sample Entropy. *Journal of the Royal Statistical Society Series B*, 38, 54-59, 1976.