

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

Jianhua Hu

Department of Biostatistics

University of North Carolina at Chapel Hill



Outline

- Introduction
- Examining clinical data and gene expression data
- Normalization and expression index estimation
- Combining estimates from different Affymetrix arrays
- Identifying important genes
- Conclusions



Introduction

- DNA microarray technology now plays an important role in many areas of biomedical research.
- Multiprobe oligonucleotide arrays have the advantage of probe redundancy.
- In our study, the two oligonucleotide array studies are explored.
 - The Michigan data set: Hu6800 platform
(20 probe pairs, 7,129 probe sets)
 - The Harvard data set: U95Av2 platform
(16 probe pairs, 12,625 probe sets)



Introduction

Research objective:

- Combining information from the two different studies.
- Identifying important genes with differential expression in normal vs. histologically-defined lung adenocarcinoma samples.
- Identifying important genes with expression related to patient survival, while incorporating the other clinical information.



Examining clinical data and gene expression data

Survival data

- Patient data from the two studies had comparable distributions of age, sex, and smoking status.
- However, there is a significant difference in survival.
- An indicator variable is created to account for an institution effect.

Examining clinical data and gene expression data

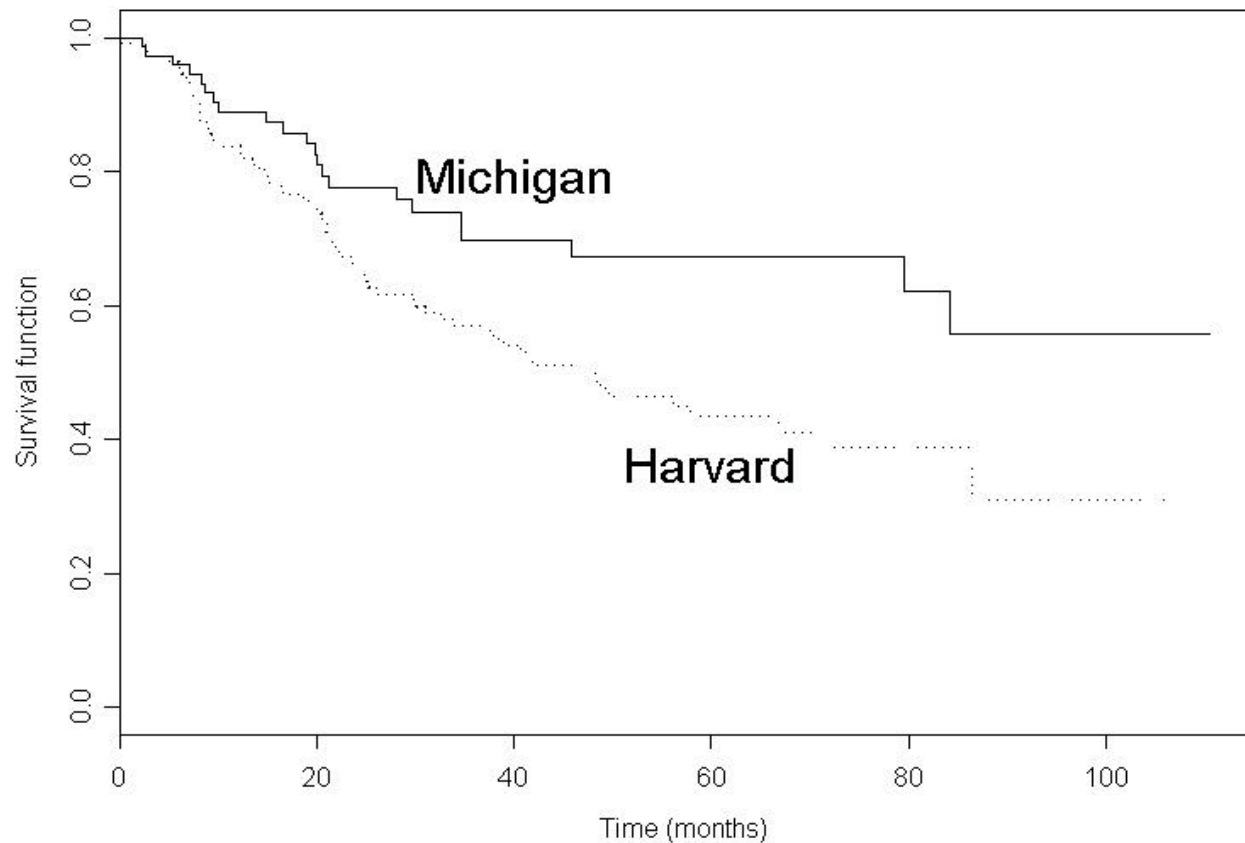


Figure 1: Estimated Kaplan-Meier survival curves.

Examining clinical data and gene expression data

Gene expression data

➤ Array outliers in Michigan data

A large round dark spot is contained at the center of the chip, e.g., L88.

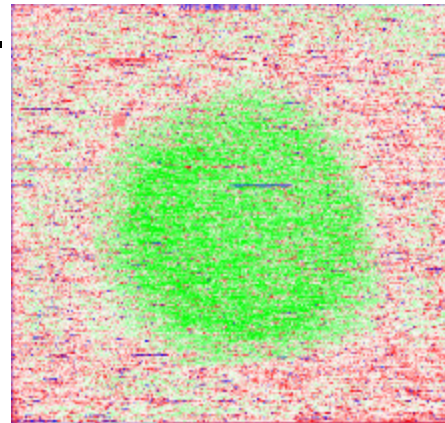


Figure 2: Green and red indicates log-expression levels below and above the median for the chip.

A large number of extremely bright outliers are contained in some arrays, e.g., L22



Examining clinical data and gene expression data

Gene expression data



Array outliers in Harvard data

Two outlier chips were detected and removed. The most recently dated run among the samples with 48 replicate arrays are kept.



Final data set contains 229 samples

143 from Harvard with 17 normal samples.
86 from Michigan with 10 normal samples.



Normalization and expression index estimation

Normalization

- Microarray normalization is important to remove sources of systematic variation in expression estimates.
- A simple linear normalization is chosen, using a synthetic “median array” as a reference.



Normalization and expression index estimation

Expression index estimation

- The term "expression index" describes a statistic used to represent an expression level for a gene.
- A multiplicative model (Li and Wong 2001a) is feasible and popular.
- The Li-Wong reduced model (LWR) using the SVD technique (Hu, Wright and Zou 2003) is performed.



Combining data from different affymetrix arrays

- A list of common probe sets representing the same gene between the two different array platforms is available at the dChip website.
- There are 5,987 probe set pairs representing the same genes across the two studies.
- The expression levels of the genes in these two chip types are not directly comparable.
- A technique for assigning weights to each expression index in the two data sets is used.

Combining data from different affymetrix arrays

➤ An important concept involved in our approach is entropy, which is defined for a continuous density $f(x)$ as $\int f(x) \log f(x) dx$.

➤ We define “fraction of eigenintensity” as $p_j = \frac{s_j^2}{\sum_{j=1}^J s_j^2}$

where J is the number of probes and s_j denotes the j th eigenvalue from the SVD decomposition.

➤ 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 $0 \leq e \leq 1$.



Combining data from different affymetrix arrays

- Assuming that the LWR is the true model from which the underlying expression index can be estimated.
- The randomness of the residual matrix can be judged by the distribution of its eigenvalues, quantified by the entropy.
- The data that better fits the model should have a higher entropy.
- To avoid one source of bias in the SVD, in each study the expression intensity matrix of each gene was standardized to a mean of 0 and a variance of 1.

Combining data from different affymetrix arrays

- Overall, the Harvard data appears much better, with residual entropies centered around 0.9, while those from Michigan are widely spread from 0 to 1.

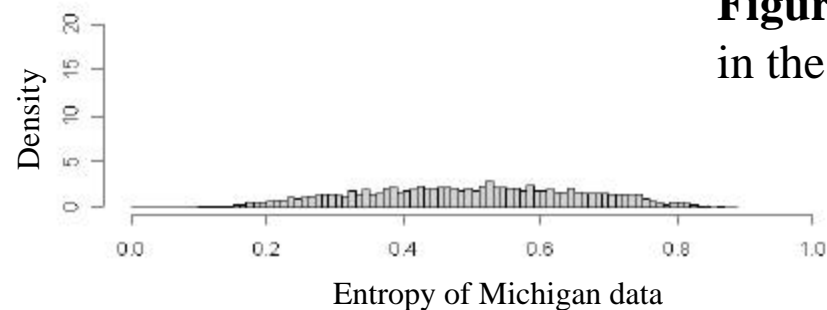
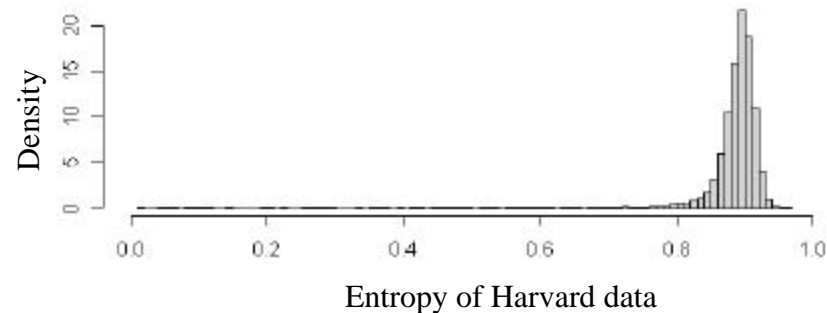


Figure 3: Distributions of entropies in the Harvard and Michigan studies.



Combining data from different affymetrix arrays

- For each gene, the two entropy values (Harvard and Michigan) were then standardized to reach a sum of 1.
- Within each study the appropriate weight was multiplied by the expression index to obtain a new entropy-weighted expression index.
- A larger weight is assigned to the model-based expression index estimate in the study that has higher entropy in the residuals for the specific gene.



Combining data from different affymetrix arrays

- To assess the performance of the entropy weighting strategy in identifying differentially expressed genes in normal vs. cancer samples, we used the false discovery rate (FDR) as a comparison criterion.
- FDR is defined as the expected proportion of false rejections (truly null) among the rejected hypotheses.
- The permutation procedures (essentially as implemented in the software SAM) is followed to estimate the FDR by using ordinary t-test statistics in normal vs. cancer samples, based on 5,000 permutations.

Combining data from different affymetrix arrays

- The weighted data yielded a lower FDR level than the unweighted one.

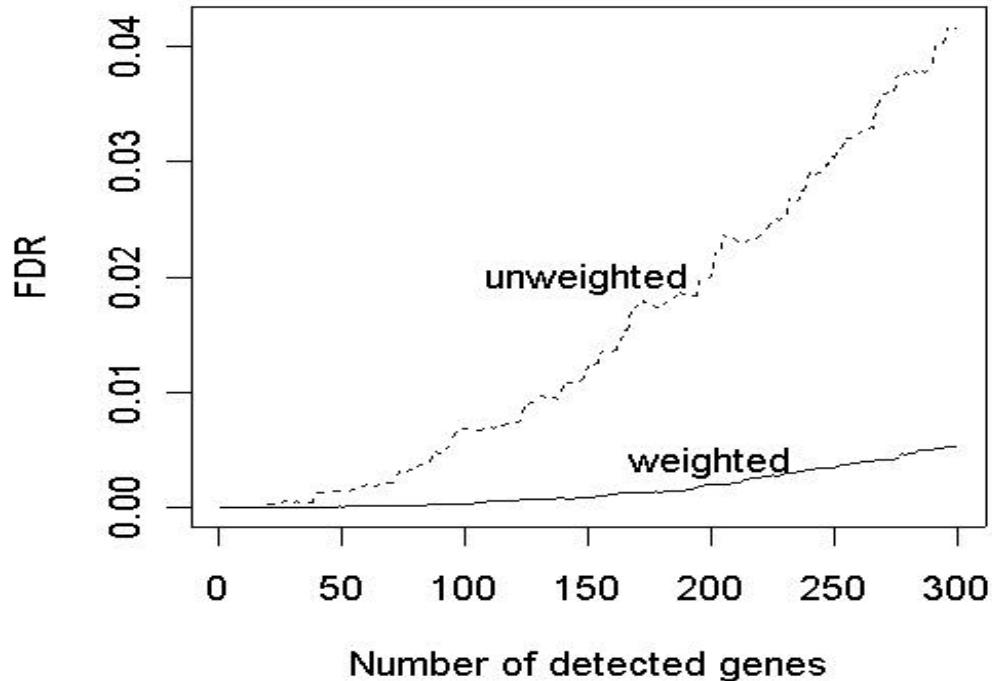


Figure 4: Comparison of FDRs between weighted and unweighted expression data.



Identifying important genes

Weighted T-Test analysis of survival data (WTT method)

- A major goal is to combine the gene expression data with the patient survival data.
- To find those genes related to the patients' survival, the clinical information needs to be taken into account, e.g., tumor stage, smoking history, sex.



Identifying important genes

The WTT method

- The Cox proportional hazards model may be applicable and amenable to entropy-weighted analysis.
- However, we devised another simple, novel approach to combine inferences of differential expression and effects of expression on survival.

Identifying important genes

The WTT method

- For the i th sample with a covariate vector Z_i , the Cox proportional hazards model is given by $l(t|Z_i) = l_0(t) \exp(\mathbf{b}^T Z_i)$
- For the i th sample, the survival function is given by $S(t | Z_i) = \exp\{-\Lambda_0(t) \exp(\mathbf{b}^T Z_i)\}$

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

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



Identifying important genes

The WTT method

- The predicted survival curve for each sample based on only the clinical information was constructed, from which the median survival time can be estimated,

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

- An averaged median survival time is assigned to those samples with missing survival information.
- m_i is determined by the covariate Z_i , which circumvents potential bias.

Identifying important genes

The WTT method

- The weights are calculated that are proportional to m_i , for each cancer patient accordingly, $w_i = \frac{m_i}{\sum_{i=1}^n m_i} \times n$
- For the normal samples, unit weights were assigned because they were controls and were all alive at the end of the study.
- With the survival-weighted expression data, we conducted a two-sample t-test for each gene (WTT) to differentiate the normal vs. cancer patients.



Identifying important genes

The WTT method

- We examined the difference between the t-test statistics after and before the survival-weight adjustment, i.e., $d_k = t_{after} - t_{before}$, for the k th gene, $k=1, \dots, 5,987$.
- We have shown that d has expectation zero for genes with no effect on survival, regardless of whether they are differentially expressed in normal vs. cancer samples.



Identifying important genes

The WTT method

- 5,000 permutations are performed. Let $d_{(k)}$ denote the ordered d_k in each permutation, the averaged order statistics, $d_{(k)}$, can be calculated.
- 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.



Identifying important genes

The WTT method

- To accommodate the multiple testing issue, we applied the FDR criterion and identified the 12 genes most significantly related to survival, while controlling for the FDR at 0.05.
- The statistical significance can be measured by p-values obtained from the permutation procedure.
- We found an intriguing number of sex-specific genes. Some other genes, including ribosomal proteins, have appeared in other cancer studies.

Identifying important genes

The WTT method

Table 2: List of 12 most important genes related to survival. Ordered from most significant to least using SAM.

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



Identifying important genes

Differentiating between normal and cancer samples

- The WTT method can be used to identify important genes differentiating the normal vs. cancer groups.
- Genes found to have significant results under both tests may be of particular interest.
- The 10 most significant genes with the smallest sums of ranks of $|d_{(k)} - d_{(k)}|$ across the two t-test statistics are identified.

Identifying important genes

Differentiating between normal and cancer samples

Table 3: The 10 most significant genes differentiating the two 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, folylpolygammagl-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



Conclusions

- We imposed a SVD entropy weight on the expression of each gene to combine different expression data.
- The approach of using residual entropy to judge the quality of expression estimates can be applied in a much more general context.
- To identify important genes having significant impact on patient survival, the WTT method is proposed. It also can be extended to more general situations, including comparisons of multiple groups.



Acknowledgements

- Guosheng Yin, Jeffrey S. Morris, Li Zhang
M.D. Anderson Cancer Center, Houston
- Fred A. Wright
UNC-Chapel Hill, Department of Biostatistics
- We thank Kevin Coombes and Fei Zou for helpful discussions.