Published in final edited form as:

J Biopharm Stat. 2008 ; 18(5): 841–852. doi:10.1080/10543400802277967.

A Predictive Risk Probability Approach for Microarray Data with Survival as an Endpoint

Dung-Tsa Chen^{1,*}, Michael J. Schell¹, James J. Chen^{2,3}, William J. Fulp¹, Steven Eschrich⁴, and Timothy Yeatman⁵

¹ Biostatistics Division, University of South Florida, 12902 Magnolia Drive, Tampa, FL 33612, USA

⁴ Biomedical Informatics, University of South Florida, 12902 Magnolia Drive, Tampa, FL 33612, USA

⁵ Surgery, Moffitt Cancer Center & Research Institute, University of South Florida, 12902 Magnolia Drive, Tampa, FL 33612, USA

² Division of Personalized Nutrition and Medicine, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR 72079, USA

³ Biostatistics Center and Department of Public Health, China Medical University, Taichung, 40402, Taiwan

Abstract

Gene expression profiling has played an important role in cancer risk classification and has shown promising results. Since gene expression profiling often involves determination of a set of top rank genes for analysis, it is important to evaluate how modeling performance varies with the number of selected top ranked genes incorporated in the model. We used a colon data set collected at Moffitt Cancer Center, as an example of the study, and ranked genes based on the univariate Cox proportional hazards model. A set of top ranked genes was selected for evaluation. The selection was done by choosing the top k ranked genes for $k=1$ to 12,500. An analysis indicated a considerable variation of classification outcomes when the number of top ranked genes was changed. We developed a predictive risk probability approach to accommodate this variation by identifying a range number of top ranked genes. For each number of top ranked genes, the procedure classifies each patient as having high risk (score = 1) or low risk (score = 0). The categorizations are then averaged, giving a risk score between 0 and 1, thus providing a ranking for the patient's need for further treatment. This approach was applied to the colon data set and demonstrated the strength of this approach by three criteria: First, a univariate Cox proportional hazards model showed a highly statistically significant level (log-rank χ^2 statistics=110 with p value $< 10^{-16}$) for the predictive risk probability classification. Second, the survival tree model used the risk probability to partition patients into five risk groups showing a good separation of survival curves (log-rank χ^2 statistics=215). In addition, utilization of the risk group status identified a small set of risk genes which may be practical for biological validation. Third, analysis of re-sampling the risk probability suggested the variation pattern of the log-rank χ^2 in the colon cancer dataset was unlikely caused by chance.

Keywords

Cancer risk classification; dimension reduction; log-rank test; survival tree model; top ranked genes

*Address correspondence to: Dung-Tsa Chen, Biostatistics Division, Moffitt Cancer Center & Research Institute, University of South Florida, 12902 Magnolia Drive, Tampa, FL 33612, USA; Email: E-mail: Dung-Tsa.Chen@moffitt.org.

The views presented in this paper are those of the authors and do not necessarily represent those of the U.S. Food and Drug Administration.

Introduction

Assessment of cancer risk in high-risk individuals and prognosis in patients are critical steps in reducing cancer-related death and improving a patient's quality of life through better assignment of specific therapies. Cancer risk classification provides an important approach for assessing risk and prognosis. Recently, gene expression profiling has been integrated in cancer risk classification development, and demonstrated markedly improved prediction results as compared to TNM staging alone (Eschrich et al., 2005; Huang et al., 2003; Sorlie et al., 2003; van't Veer, Paik and Hayes, 2005; van de Vijver et al., 2002; West et al., 2001).

Various statistical methods to correlate gene expression with time-to-event endpoint for the cancer risk classification have been developed, such as unsupervised clustering techniques with the log-rank test, partial least squares, or principal component analysis (PCA) with the Cox proportional hazards model (Bair and Tibshirani, 2004; Nguyen and Rocke, 2002; Park, Tian and Kohane, 2002; Rosenwald et al., 2002). Basically, these methods use various algorithms to reduce gene expression data from a high-dimensional space to a manageable low-dimensional space for the follow-up statistical survival analysis. While these methods have their applications in medical research, it remains unclear how many genes should be chosen for analysis. Although criteria such as p value, false discovery rate (Benjamini and Hochberg, 1995), fold change, or the combination, have widely been used to select a set of genes for profiling, the number of genes selected is subjective.

In this paper, we conducted a comprehensive study to evaluate the effect of the selection of top ranked genes using a colon data set collected at Moffitt Cancer Center. Genes were ranked based on the univariate Cox proportional hazards model; and a set of top ranked genes was selected for evaluation. The selection was done by choosing the top k ranked genes for $k=1$ to 12,500. The symbol, N_k , represents the top k ranked genes selected for the follow-up statistical analysis. Since the evaluation is computationally intensive, we limited the study to a PCA approach. Considerable variation among classification outcomes was found when the number of top ranked genes changed. We developed a predictive risk probability approach to accommodate this variation.

Methods

Colon Cancer Dataset at the Moffitt Cancer Center

The colon cancer dataset was collected from 205 colon cancer patients at the Moffitt Cancer Center. One primary study aim was to develop a gene profiling signature associated with overall patient survival. The Affymetrix U133 plus 2 gene chip was used to obtain whole genomic gene expression for these patients with one chip per patient. There are 54,675 probe sets per gene chip. Expression data was pre-processed using the RMA method (Irizarry et al., 2003). Overall survival time ranged from 0.92 to 142 months with a 58% censoring rate (Table 1). Most patients (64%) had Dukes stage B or C, 16% patients had Dukes stage A, and 20% patients had Dukes stage D (Table 2).

PCA Algorithm

Principal component analysis (PCA) is a powerful tool to analyze high-dimensional data by reducing data dimensions into a manageable space. Data reduction of PCA is performed by a set of uncorrelated principal components where each principal component is a linear combination of individual gene expressions. A set of selected components which maximize the total variation of gene expressions is then used as predictors for the follow-up statistical analysis. We select PCA for evaluation because it is easy for implementation and less computationally intensive than other algorithms. Since a principal component with a large

variation may not correlate strongly with survival time, we consider the following steps to overcome this issue.

1. **Gene selection:** We used the univariate Cox proportional hazards model to calculate the p value for each gene and then order genes with the smallest p value as the highest rank. This ranking system is likely to help select top ranked genes strongly associated with survival time.
2. **Pre-specification of percentage cutoff of total variation:** Expression data of the selected k top ranked genes is transformed by PCA to a set of principal components (data reduction). The percentage cutoff of total variation (e.g., 90%) is then used to choose the first c principal components as independent variables for survival data analysis.
3. **Backward elimination:** The c principal components are uncorrelated and explain a certain percentage of total variation, However, not all of these principal components are strongly associated with survival endpoint. The backward selection procedure allows elimination of less relevant principal components to build a more parsimonious final model. Here we use Akaike information criterion (AIC) to determine whether a single principal component should be dropped from the model (Hastie and Pregibon, 1992; Venables and Ripley, 2002). Specifically, at each step, AIC is computed for the full model and for the sub-models with one principal component deleted (e.g., If there are k principal components in the full model, there will be k AICs for the sub-models which include only $k-1$ principal components, as well as one AIC for the full model). The calculated AICs are then ranked from the smallest to the largest. If AIC of the full model is not the smallest, the principal component corresponding to the sub-model which does not include this principal component and have the smallest AIC is dropped. The process is repeated until the full model reaches the smallest AIC.
4. This final model has a covariate form in the log scale of $\mu_l = \sum_{i \in A} \alpha_i \times PC_{i,l}$, where μ_l is a composite risk score for subject l to reflect the relative risk (e^{μ_l}), $PC_{i,l}$ represents the i^{th} principal component score for the subject l and α_i is the corresponding coefficient in the Cox proportional hazards model, and A is the set of indices of the remaining principal components after backward elimination.
5. **Classification of patient risk:** The model from the step 4 calculates the composite risk score (μ_l) for subject l for $l=1, \dots, n$ where n is the total number of subjects. A positive μ_l indicates a relative risk greater than 1 and a negative one means a relative risk less than 1. We use $\mu_l=0$ as the cutoff to classify subjects (i.e., $\mu_l > 0$ as high risk and $\mu_l \leq 0$ as low risk).
6. **Comparison of survival curves:** Classification from the step 5 splits subjects into two risk groups (high and low risk). Kaplan-Meier curve and the log-rank test are used to examine the separation magnitude of survival curves. This process transforms the high dimensional gene expression variables (continuous variables) into a risk status (a dichotomous variable: high and low risk) for each patient which allows for the use of Kaplan-Meier curve and the log-rank test in the survival analysis.

Variation of log-rank χ^2 statistics

We examined the impact of different numbers of top ranked genes on the proposed PCA algorithm. We limited the number of top ranked genes from 1 to 12,500 genes because the algorithm became unstable beyond 12,500. Figure 1 is the results of the log-rank χ^2 statistics over the 12,500 different numbers of top ranked genes. The results showed a large variation

even for a small change of the number of top ranked genes. Such variation raised a serious concern of the practical use of the number of top ranked genes on microarray data analysis.

Predictive risk probability

We develop a predictive risk probability approach to account for this variation. Specifically, for each number of top ranked genes, N_k , the procedure classifies each patient as having high risk (score = 1) or low risk (score = 0). We repeat this step for a wide range number of top ranked genes (i.e., N_k for $k=k_1$ to k_2 , denoted by $N_{k_1}^{k_2}$) to obtain a set of risk scores for each patient. Using the principle of majority vote, this approach averages these classification results to generate a predictive risk probability which can be used to rank the patient's risk and to provide the need for further treatment or other intervention.

For example, we analyzed the colon cancer data using 12,500 different N_k . Each patient thus had 12,500 classification risk outcomes (high or low risk). We averaged these classification risk outcomes being at high-risk as the predictive risk probability. Suppose patient A is classified as high risk for the first 11,000 N_k ($N_1, \dots, N_{11,000}$), and as low risk at the last 1,500 N_k ($N_{11,001}, \dots, N_{12,500}$), the predictive risk probability is 88% at high risk for this patient.

Evaluation

To evaluate the predictive risk probability, we performed a series of analyses to examine the approach, including the Cox proportional hazards model, survival tree model, and re-sampling scheme to examine the variation pattern.

Cox proportional hazards model

We studied the association of the predictive risk probability with survival time and the p value of univariate Cox model, as well as its effect on the univariate Cox model.

Association of the risk probability with survival time—The predictive risk probability can be used as a predictor for survival data analysis. In this study, a Cox proportional hazards model was run with this predictive risk probability as the only independent variable. The hazard coefficient was 4.28 with a high log-rank score ($\chi^2=110$, $p<10^{-16}$). Note that the large value of the hazard coefficient indicated that a higher risk probability generated a higher hazard rate. The result showed that the predictive risk probability was an important predictor of survival time.

Correlation of the risk probability with p value of the univariate Cox model—Since the risk probability was highly associated with survival, it is likely that the risk probability is also correlated with significant genes (i.e., genes with a significant p value in the univariate Cox model). We calculated Pearson correlation of the risk probability and expression for each gene. Result showed a moderate correlation between the risk probability and gene expressions (−0.4 to 0.4). We plotted the correlation versus the p values of the univariate Cox model in Figure 2. The plot showed an interesting nonlinear pattern between the correlation and the p values. For a weak correlation between −0.1 to 0.1, most p values were larger than 0.01. When correlation becomes stronger (greater than 0.35 or less than −0.35), the p values were all less than 0.01.

Survival tree model

We used a survival tree model (Therneau and Atkinson, 2006) to partition patients into groups based on the risk probability. Categorization of the risk probability by the tree model was to identify a set of risk genes.

Grouping the predictive risk probability—The survival tree model was utilized to classify patients into five risk groups using the risk probability as the only covariate: risk probability <0.06 (n=34), 0.06–0.52 (n=61), 0.52–0.91 (n=40), 0.91–0.98 (n=29), and >0.98 (n=41) (Figure 3). Sample size ranged from 29 to 61 among the five risk groups. The five survival curves were well-separated with a log-rank statistics of 215 and a p value less than 0.00001. Pair-wise comparison (10 comparisons) among the 5 risk groups showed most p values were small (eight had p values less than 0.005 and two had a p value of 0.01). Compared to the Dukes staging system, the five risk groups derived from the risk probability showed a positive association (Table 3). Specifically, patients with stage A or B tended to have a risk probability below 0.50 (the chance is 2:1). Similarly, for patients with stage C or D, their risk probability tended to be high (>0.5 with a chance of 2:1).

Identification of risk genes—It is important to find a subset of genes as risk genes for biological validation. While the risk probability proved to be a promising predictor for this survival data, it used a large number of genes which limited its application for validation. However, by utilization of the risk group status (i.e., categorization of the risk probability), we could identify a subset of risk genes to differentiate risk groups (e.g., high versus low risk groups). Here we illustrate this concept by comparing two extreme risk groups (the lowest risk group (n=34): risk probability<0.06 versus the highest risk (n=41) group: risk probability>0.98). Since this strategy converts survival data into two (or multiple) risk groups, standard statistical methods for group comparison (e.g., Significance Analysis of Microarrays (SAM) (Tusher, Tibshirani and Chu, 2001)) can be used for analysis. Here we used SAM method to choose a set of risk genes by controlling FDR<0.01 and a fold change>2 and employed the support vector machine (SVM) (Fan, PH Chen and Lin, 2005) to build a predictive model based on the selected risk genes. The whole process was evaluated using leave-one-out cross validation (LOOCV) to assess accuracy. Specifically, at each loop of LOOCV, data of one subject was held out as test set (n=1) and the rest of data was used as training data (n=74). SAM and SVM were used for gene selection and construction of a predictive model in the training set. The test set was then plugged in to the predictive model to see if the classification status matched the true risk status (high or low risk). Evaluation result by LOOCV showed a reasonably high accuracy rate, 88%, indicating the adequacy of the methods used for the data. The final model, which used the whole data, identified 60 probe sets with significantly differential expression. While LOOCV is a strategy to validate methods employed for data analysis, gene lists generated at each LOOCV could also be used to check sensitivity of significant genes when data is changed. By comparing to the results from LOOCV which generated 75 gene lists, 59 probe sets (out of 60: 98%) were picked up at least 40 times (out of 75 times: 53%) and 43 probe sets (72%) were always identified at each LOOCV (Table 4). This observation suggests the robustness of these probe sets. Plots of the first principal component based on these 60 probe sets showed a good separation of the two risk groups in Figure 4.

Examination of the variation pattern by re-sampling the risk probability

Figure 1 had shown a large variation of the log-rank χ^2 statistics for $N_1, \dots, N_{12,500}$. The lowest smooth curve was plotted to show that the χ^2 statistics increased as the number of top ranked genes increased from 1 to 1,000 ($N_1, \dots, N_{1000}$). The χ^2 statistics reached peak and became stable when the number of top ranked genes was between 1,000 and 2,500 ($N_{1000}, \dots, N_{2500}$). As the number of top ranked genes increased (> 2,500), the χ^2 statistics started to gradually decrease. One may wonder if this variation pattern was due to random chance. To address this concern, we took a supportive approach to examine whether the variation pattern is different between the case of high association of the risk probability with survival time and the case with weak association. Specifically, we used the calculated predictive risk probability to re-sample survival time. As shown previously, the risk probability was very significant in the Cox model

($\chi^2=110$) and the survival tree model classified the risk probability into 5 risk groups. We used these five risk groups as the re-sampling framework. We partitioned each risk group into 5 sub-groups. The partition was done using random sampling without replacement. Survival time and censor status in each sub-group were swapped with another sub-group in the other risk group while retaining gene expression data. Since the 5 risk groups had unequal sample size, exchange of survival information between two sub-groups was done as follows. Suppose sub-group A from one risk group has k samples and sub-group B (for trade) from another risk group has $k+m$ samples, we randomly choose k samples from the sub-group B to be swapped with the k samples from the sub-group A. This re-sampling scheme likely breaks down the effect of the risk probability. If there was no effect by the risk probability, we should see a similar variation pattern like Figure 1 when analyzing the re-sampling data. Results of two re-sampling datasets showed an increasing pattern of χ^2 values by the lowess smooth curve as N_k increased in Figure 5, different from the variation pattern in Figure 1. This observation indicated that, for the case of no effect of the risk probability, the χ^2 value continued increasing as N_k increased. On the other hand, with strong association of the risk probability with survival, the χ^2 value was likely to decrease when N_k became very large.

Discussion

We have studied the impact of the selection of the number of top ranked genes, N_k , on microarray data analysis with survival endpoint for a 205 patient colon cancer study. We observed considerable variation of log-rank χ^2 statistics as the number of top ranked genes varied. To account for such variation, we have developed a predictive risk probability approach. Instead of making a decision from simply some set number of top ranked genes, this predictive risk probability approach identifies a range number of top ranked genes. From each number of top ranked genes in the range the analysis identifies each patient as high risk (score = 1) or low risk (score = 0). The categorizations are then averaged, giving a high risk score between 0 and 1, thus providing a ranking for the patient's need for further treatment or other intervention. We have used several criteria to evaluate this approach and demonstrated its strength: (a) A univariate Cox proportional hazards model showed a highly statistically significant level ($p < 10^{-16}$) for the predictive risk probability. The risk probability was also correlated with significant genes; (b) Using the risk probability as the only covariate, survival tree model effectively grouped patients into five groups with good separation of survival curves (log-rank statistics=215). The five survival curves were in order of the predictive risk probability (A higher risk probability gives a shorter survival). In addition, utilization of the risk group status identified a small set of risk genes which makes practical for biological validation; and (d) Analysis of re-sampling the risk probability indicated the variation pattern of the log-rank χ^2 for the case of no risk probability effect was different from the one when the risk probability was strongly associated with survival.

While the predictive risk probability approach demonstrates its usefulness, we would like to point out several issues for improvement in future study: (1) We have evaluated the predictive risk probability using a wide range number of top ranked genes from 1 to 12,500 ($N_1^{12,500}$). The result may be enhanced by using an appropriate range number of top ranked genes. A rigorous statistical method to find an optimal range will greatly improve the risk probability approach; (2) The log-rank χ^2 statistics may be affected by N_k and the number of principal components in the final model after the backward elimination step. It will be worthwhile to study their relationship; (3) This study uses p values from the univariate Cox proportional hazards model for gene ranking. Other criteria or survival methods could be used to define the top ranked gene list, too, such as the combination of p value with fold change and parametric exponential survival model; (4) The effect of the number of top ranked genes is limited to the PCA approach in gene expression profiling. Generalization to other data reduction approaches

(e.g., PLS and clustering) and other datasets would be an interesting future research; (5) The predictive risk probability approach is computationally intensive. When the number of top ranked genes becomes very large, PCA and backward deletion steps require substantial amount of computation time to finish the job. The problem could be improved by using parallel computing in cluster machine.

Acknowledgments

This work was supported by grants from the National Institute of Health (1P30CA76292-07 and R01 CA098522). The authors would like to acknowledge the use of the services provided by Research Computing, University of South Florida.

References

- Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* 2004;2:E108. [PubMed: 15094809]
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological* 1995;57:289–300.
- Eschrich S, Yang I, Bloom G, et al. Molecular staging for survival prediction of colorectal cancer patients. *J Clin Oncol* 2005;23:3526–3535. [PubMed: 15908663]
- Fan R, PH Chen, Lin C. Working Set Selection Using the Second Order Information for Training SVM. *Journal of Machine Learning Research* 2005;6:1889–1918.
- Hastie, TJ.; Pregibon, D. Generalized linear models. Chapter 6 of *Statistical Models in S*. Chambers, JM.; Hastie, TJ., editors. Wadsworth & Brooks/Cole; 1992.
- Huang ES, Black EP, Dressman H, West M, Nevins JR. Gene expression phenotypes of oncogenic signaling pathways. *Cell Cycle* 2003;2:415–417. [PubMed: 12963829]
- Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003;31:e15. [PubMed: 12582260]
- Nguyen DV, Rocke DM. Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics* 2002;18:1625–1632. [PubMed: 12490447]
- Park PJ, Tian L, Kohane IS. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics* 2002;18(Suppl 1):S120–127. [PubMed: 12169539]
- Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med* 2002;346:1937–1947. [PubMed: 12075054]
- Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* 2003;100:8418–8423. [PubMed: 12829800]
- Therneau TM, Atkinson B. rpart: Recursive Partitioning. R Package. 2006
- Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 2001;98:5116–5121. [PubMed: 11309499]
- van't Veer LJ, Paik S, Hayes DF. Gene expression profiling of breast cancer: a new tumor marker. *J Clin Oncol* 2005;23:1631–1635. [PubMed: 15755970]
- van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002;347:1999–2009. [PubMed: 12490681]
- Venables, WN.; Ripley, BD. *Modern Applied Statistics with S*. Vol. 4. New York: Springer; 2002.
- West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A* 2001;98:11462–11467. [PubMed: 11562467]

Figure 1. Distribution of log-rank χ^2 statistics for the number of top ranked genes from 1 to 12,500. The lowest smooth curve was plotted to display variation of the χ^2 statistics.

Figure 2.
A nonlinear relationship between the predictive risk probability and gene expression.

(a)

Survival Tree of Predictive Risk Probability (PRP)

(b)

Log-rank test=215 ($p < 0.00001$) **Figure 3.**

(a) Survival tree analysis using the predictive risk probability as the predictor; (b) Kaplan-Meier survival curves of the 5 risk groups from the tree model.

Figure 4. Distribution of the first principal component based on the 60 probe sets between the low (L) and high (H) risk groups.

Figure 5. Distribution of log-rank χ^2 statistics based on two re-sampling datasets (breaking down the risk probability effect) for the number of top ranked genes from 1 to 12,500. The lowest smooth curve was plotted to display variation of the χ^2 statistics.

Table 1
Distribution of overall survival by year of the colon cancer dataset.

Distribution of Overall Survival													
year	1	2	3	4	5	6	7	8	9	10	11	12	Total
Number of alive	4	8	13	13	25	12	13	10	6	7	6	1	118
Number of dead	22	17	19	10	8	7	2	1	0	0	0	1	87
Total	26	25	32	23	33	19	15	11	6	7	6	2	205
Censoring rate=58%													

Table 2
Distribution of Dukes stage of the colon cancer dataset.

Distribution of Dukes Stage				
	A	B	C	D
Frequency (%)	32 (16%)	66 (32%)	65 (32%)	42 (20%)
Total				205

Table 3
Comparison of the five prognostic groups derived from the risk probability and the Dukes stage system.

	Risk Probability					
	<0.06	0.06–0.51	0.51–0.91	0.91–0.98	>0.98	
Dukes Stage	A	11	13	6	2	0
	B	12	29	9	9	7
	C	8	15	15	12	15
	D	3	4	10	6	19

Distribution of the 60 probe sets (derived from the comparison of the high and low risk groups) being classified as significant genes in the 75 loops of the leave-one-out cross validation (LOOCV).

Table 4

Frequency of being classified as significant gene	37	40	44	45	58	59	61	63	66	67	70	72	74	75
Number of probe sets	1	1	1	2	2	1	1	1	1	2	2	1	1	43