Bioinformatics Advance Access published June 16, 2005

BIOINFORMATICS

Use of Extreme Patient Samples for Outcome Prediction from Gene Expression Data

Huiqing Liu, Jinyan Li and Limsoon Wong*

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613

ABSTRACT

Motivation: Patient outcome prediction using microarray technologies is an important application in bioinformatics. Based on patients' genotypic microarray data, predictions are made to estimate patients' survival time and their risk of tumor metastasis or recurrence. So, accurate prediction can potentially help to provide better treatment for patients.

Results: We present a new computational method for patient outcome prediction. In the training phase of this method, we make use of two types of extreme patient samples: *short-term survivors* who got an unfavorable outcome within a short period and *long-term survivors* who were maintaining a favorable outcome after a long follow-up time. These extreme training samples yield a clear platform for us to identify relevant genes whose expression is closely related to the outcome. The selected extreme samples and the relevant genes are then integrated by a support vector machine to build a prediction model, by which each validation sample is assigned a risk score that falls into one of special pre-defined risk groups. We apply this method to several public data sets. In most cases, patients in high and low risk groups stratified by our method have clearly distinguishable outcome status as seen in their Kaplan-Meier curves. We also show that the idea of selecting only extreme patient samples for training is effective for improving the prediction accuracy when different gene selection methods are used.

Contact: huiqing@i2r.a-star.edu.sg

Supplementary information: http://research.i2r.a-star.edu.sg/huiqing/supplementaldata/survival/survival.html

INTRODUCTION

Currently, the risk of a cancer patient is mostly measured by various clinical factors, such as size of the original tumor, extent of local invasion, spread to distant organs, and so on. However, in many cases, patients with a similar clinical diagnosis may have different responses to the same treatment. For example, for patients suffering from acute myeloid leukemia, the most common acute leukemia in adults, chemotherapy can induce a complete remission in 70-80% of younger patients, but many of them have a relapse and die of their disease (Bullinger *et al*., 2004). Though a stem-cell transplantation approach can prevent relapse of this disease, this approach is associated with a high treatment-related mortality (Bullinger *et al*., 2004). Therefore, accurate outcome prediction methods are needed to personalize the treatment plan for each individual patient. Thus, improper treatment and subsequent severe sufferings for patients (e.g. decline in IQ, hormonal deficiency problems) or inefficient treatment that causes relapse can be avoided.

Microarray technology enables monitoring of disease progression and prediction of patient outcome at the molecular level. A few previous studies have shown promising results for survival prediction from gene expression profiles and clinical data for certain diseases (Rosenwald *et al*., 2002; Beer *et al*., 2002; van de Vijver *et al*., 2002; Yeoh *et al*., 2002; Bullinger *et al*., 2004). These studies have demonstrated to be useful for optimizing treatment plans for individual patient, and also have recommended candidate genes that may be useful for developing innovative therapies and generating opportunities for drug discovery.

In early approaches proposed for outcome prediction from gene expression profiles, the traditional Cox proportional hazards model (Cox, 1972; Lunn and McNeil, 1995) is usually used to select genes. By this model, genes most related to survival are identified by a univariate Cox analysis, and a risk score is then defined as a linear weighted combination of the expression values of the identified genes (Beer *et al*., 2002; Rosenwald *et al*., 2002). Recently, machine learning technologies are involved. For example, Ando and Katayama (2002) have proposed a fuzzy neural network system to predict survival of patients using gene expression profiles as input; Park *et al*. (2002) have developed a method to co-relate gene expression data to patient survival time using a partial least squares regression technique; and Shipp *et al*. (2002) have employed a weighted voting algorithm to identify cured versus fatal for outcome prediction of diffuse large B-cell lymphoma (data set of Rosenwald *et al*., 2002). In a more recent work (LeBlanc *et al*., 2003), a gene index technique has been introduced to identify the associations between gene expression levels and patient outcome. The core idea of their method is to combine the correlation between genes with patient outcome as well as class membership for the ranking.

*to whom correspondence should be addressed

Very recently, a semi-supervised method has been proposed to make use of both clinical information and gene expression profile for outcome prediction (Bair and Tibshirani, 2004). In their method, a subset of genes whose Cox score exceeds a certain threshold are chosen, and then unsupervised learning techniques (clustering or principal components analysis) are applied to these genes to group patients into different risk categories. An important suggestion made by Bair and Tibshirani (2004) is that analysing patients with different survival rates based on gene expression data would help identify subtypes of cancer.

In this paper, we present a new computational method for outcome prediction based on gene expression profiles. Different from all previous works, our idea to form the training data is novel. Our training data consists of only two types of extreme patient samples: *short-term survivors* who got an unfavorable outcome within a short period and *long-term survivors* who were maintaining a favorable outcome after a long follow-up time. We do not consider patient samples between the two extreme cases in the training. A reason to select these extreme patient samples for training is that the sharp contrast between short-term and long-term survivors should be more informative and reliable (than those medium-term cases) for building and understanding the relation between genes and outcome. The addition of the medium-term patient samples would bring more noise and confusion signals.

To identify genes most associated with the outcome, we apply a two-phase feature selection method to the selected training data. The two-phase feature selection method combines an entropy measurement (Fayyad and Irani, 1993) and the Wilcoxon rank sum test method (Wilcoxon, 1945) for identifying those sharp discriminative genes. To construct a model for survival risk estimation, we train a linear kernel support vector machine (SVM) based on the selected training samples and the selected genes. When a patient sample is given for outcome prediction, we calculate a risk score by feeding the patient's expression profiles to the established model. Based on this score, we then assign this patient to one of pre-defined risk groups such as high risk, intermediate risk, or low risk group. Explicit threshold values to categorize different risk groups can be easily obtained based on the training results, so that outcome prediction for new patients is possible.

We apply our method to three large data sets: a data set consisting of 240 patients of diffuse large-B-cell lymphoma (Rosenwald *et al.*, 2002), a data set of 116 patients of adult acute myeloid leukemia (Bullinger *et al.*, 2004), and a data set of 295 patients of breast cancer (van de Vijver *et al.*, 2002). The corresponding Kaplan-Meier plots illustrate that the patients assigned to different risk groups based on our risk score have significantly different outcome. To further examine the idea of the training sample selection, we use a different feature selection method, called SAM (significance analysis of microarrays) (Tusher *et al.*, 2001), to find genes associated with outcome. Comparisons with results of this approach demonstrate again the effectiveness of our extreme training sample selection idea.

METHODS

We first present the new idea of selecting extreme training samples. Then we describe how to identify outcome-relevant genes from these training samples. Then, we introduce a scoring function for patients' risk estimation and outcome prediction.

Selection of extreme patient samples for training

Since our focus is on the relationship between gene expression and outcome, two types of extreme cases — *short-term survivors* who got an unfavorable outcome within a short period and *long-term survivors* who were maintaining a favorable outcome after a long follow-up time should be more valuable than those in the "medium-term" status. I.e., we do not expect reliable prediction could come out from analysing alive patients whose available follow-up time is short. So, the training data used in our method consists of only these two types of samples. This idea is different from all previous approaches that always use all training samples.

Specifically for an experimental sample T, if its follow-up time is $F(T)$ and status at the end of follow-up time is $E(T)$, then

$$T\ is \begin{cases} \text{short-term survivor,} & \text{if } F(T) < c_1 \wedge E(T) = 1 \\ \text{long-term survivor,} & \text{if } F(T) > c_2 \\ \text{others,} & \text{otherwise} \end{cases} \quad (1)$$

where, $E(T) = 1$ stands for "dead" or an unfavorable outcome, $E(T) = 0$ stands for "alive" or a favorable outcome, c_1 and c_2 are two thresholds of survival time for selecting short-term and long-term survivors, respectively. Note that long-term survivors also include those patients who died after a very long period. The two thresholds, c_1 and c_2, can vary from disease to disease and from data set to data set. Our basic guide line for the selection of c_1 and c_2 is that the selected training data should contain enough training samples for learning algorithms: generally, we require that each class should have at least 10 samples, and the total number of extreme samples should be between one fourth and one half of all available samples.

Identification of relevant genes

We propose a two-phase feature selection method to identify genes expressed differentially between short-term and long-term survivors. In the first phase, we use an entropy-based feature selection method to identify those features whose expression statistically differ between the two types of extreme patient samples. In this phase, motivated by the study of Liu and Setiono (1995) that discretization has the potential to perform feature selection among numeric features, we apply a supervised discretization algorithm (Fayyad

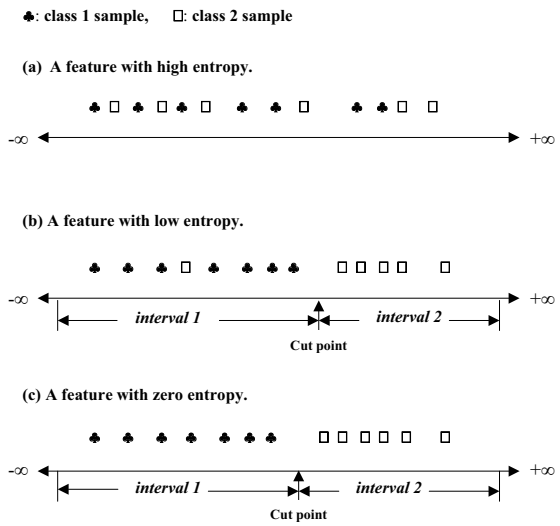


Fig. 1. We place the values of a feature on the horizontal axis. There are 13 samples in two classes, class 1 and class 2. (a) shows a feature that is a poor signal and there is no cut point can be found to distinguish samples in the different classes; (b) shows a feature that is a potentially good signal and indicates a possible cut point. (c) shows a feature that is a strongest signal and indicates a cut point — different resulting intervals contains samples of different class.

and Irani, 1993) to each of genes. This algorithm partitions a value range of a numeric feature in a way such that each of the resulting intervals contains the same class of samples as many as possible. For those features whose values are relatively randomly distributed between different classes of samples, the algorithm does not partition the value range, or, in other words, the feature can be only discretized to a single value. This means that those features do not make any significant contribution to the separation of the different classes of samples. Therefore, we discard them from our analysis. On the other hand, if a resulting value interval induced by the cut points of a feature contains only the same class of samples, then this partitioning of this feature has an entropy value of 0. This is an ideal case since the feature can clearly distinguish samples in the different classes. Figure 1 briefly illustrates the entropy measure, cut points and intervals. From our previous study (Li *et al.*, 2003), this systematic method usually discretizes less than 10% of the original features. For details of the algorithm, interested readers are referred to Fayyad and Irani (1993) and our supplementary web site.

In the second phase, we use the Wilcoxon rank sum test (Wilcoxon, 1945) to narrow down the features selected in the first phase by selecting only those more sharply discriminating features. It is a kind of non-parametric test since it is based on the rank of samples rather than distribution parameters such as mean and standard deviation. Wilcoxon rank sum test is an alternative to t-test but has several advantages such as its good tolerance to outliers and its robustness to data transformation. These characteristics make the Wilcoxon rank sum test a favorable feature selection method in gene expression profile study (Park *et al.*, 2001; Troyanskaya *et al.*, 2002). Given a gene X with its test statistical measure $w(X)$ calculated by the Wilcoxon rank sum test, if $w(X)$ falls in the interval $[c_{lower}, c_{upper}]$, where c_{lower} and c_{upper} are the lower and upper critical test values, then X is removed from further consideration. Otherwise, gene X is selected because it rejects the null hypothesis, and thus its expression values are significantly different between the two classes. In the calculation of the two critical values c_{lower} and c_{upper}, a significant level of 5% or 1% is generally used. We use 5% in this paper. A description of the method can be found at our supplementary web site.

Construction of a SVM scoring function

We propose a new scoring function to estimate the outcome for every patient. This scoring function is based on support vector machines (SVM) (Vapnik, 1995). The implementation of our SVM is by *Weka* (version 3.2), available at `http://www.cs.waikato.ac.nz/ml/weka`. In our case, the SVM regression function $G(T)$ is a linear combination of the expression values of the identified genes:

$$G(T) = \sum_i \alpha_i y_i K(T, x(i)) + b \qquad (2)$$

where the vectors $x(i)$ are the support vectors (samples), y_i are the class labels (1 and -1 used here) of $x(i)$, vector T represents a test sample, and α_i and b are numeric parameters can be determined from the training data.

We map the class label "short-term survivors" to 1 and "long-term survivors" to -1. If $G(T) > 0$, then the sample T is more likely to be a "short-term survivor". If $G(T) < 0$, then the sample T is more likely to be a "long-term survivor". To transform the output of $G(T)$ into probability-like values, we use a standard sigmoid function $S(T)$ defined as:

$$S(T) = \frac{1}{1 + e^{-G(T)}} \qquad (3)$$

So, $S(T)$ is in the range $(0, 1)$. Also note that the smaller the $S(T)$ value is, the better outcome the patient T will have. We term $S(T)$ as the *risk score* of T.

If one only categorizes patients into high risk or low risk groups, the value 0.5 is a natural cutoff for $S(T)$. In other words, if $S(T) > 0.5$ then the patient T will be assigned to high risk group; otherwise, the patient will belong to low risk group. If more than two risk groups are considered — such as high, intermediate, and low — then other cutoffs can be determined based on the risk scores of the training samples.

E.g., in training set, if most of short-term survivors have a risk score greater than r_1 and most of long-term survivors have a risk score smaller than r_2, then,

$$T\ is \begin{cases} \text{high risk}, & \text{if } S(T) > r_1 \\ \text{low risk}, & \text{if } S(T) < r_2 \\ \text{intermediate risk}, & \text{if } r_2 \leq S(T) \leq r_1 \end{cases} \quad (4)$$

In general, we require r_1¿0.5, $r_2 < 0.5$; the selection of the precise values of r_1 and r_2 can be guided by the risk scores of the training samples.

To visualize the probability of survival of all patients in different risk groups, we draw Kaplan-Meier curves (Altman, 1991) for all the groups. A point in a survival curve indicates the survival fraction (or percentage) of the patients in the group at a specific time. In this study, the Kaplan-Meier survival curves are generated by *GraphPad Prism* (`http://www.graphpad.com`). To compare the survival characteristics between different risk groups, *log-rank test* is used. The log-rank test generates a p-value testing the null hypothesis that the survival curves are not different between two groups. The meaning of p-value is that "if the null hypothesis is true, what is the probability of randomly selecting samples whose survival curves are not different from those actually obtained?". So if the p-value is small, the difference between groups is statistically significant. In this paper, we report p-value at 95% confidence interval.

EXPERIMENTS AND RESULTS

This section reports our results on three public data sets. To demonstrate the high effectiveness of our extreme sample selection method, we also report good outcome prediction results obtained by using a different feature selection method (instead of our two-phase feature selection method) to pick up important features from extreme training samples for constructing the SVM model. This feature selection method is called *SAM* (Significance Analysis of Microarrays), which is a software developed at Stanford University (`http://www-stat.stanford.edu/~tibs/SAM/`). See our supplementary web page for more information about SAM.

Diffuse large-B-cell lymphoma

Survival after chemotherapy for diffuse large-B-cell lymphoma (DLBCL) patients has been previously studied by Rosenwald *et al* (2002) based on gene expression profiling and a Cox proportional hazards model. In that study, expression profiles of biopsy samples from 240 patients are used. The data include a preliminary group consisting of 160 patients and a validation group of 80 patients, each of them is described by 7399 microarray features.

We set c_1=1 year and c_2=8 years in Formula (1) to select short-term and long-term survivors from the preliminary 160-patient group. There are in total 47 short-term survivors and 26 long-term survivors. So, our training set consists of only 73 samples. From these 73 extreme patient samples,

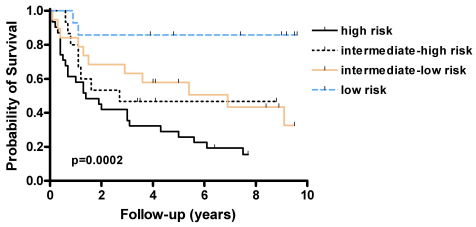


Fig. 2. Kaplan-Meier plots illustrate the estimation of overall survival among four different patient risk groups for a DLBCL study. A tick mark on the plot indicates that one sample is censored at the corresponding time.

we identified 84 features that are related to patient survival status by using our two-phase feature filtering method. Interestingly, some of the selected genes are also listed in the Table 2 of (Rosenwald *et al.*, 2002), where significant survival-associated genes are reported and previously studied. The common ones include *AA805575* (GenBank accession number) in *germinal-center B-cell signature*, *X00452* and *M20430* in *MHC class II signature*, and *D87071* in *lymph-node signature*. Some other top-ranked genes (with smaller entropy value) in our list also have clear gene signatures. For example, *BC012161*, *AF061729* and *U34683* are in *proliferation signature*, *BF129543* is in *germinal-center B-cell signature*, and *K01144* and *M16276* are in *MHC class II signature*.

We constructed a good SVM model based on the 73 extreme training samples and the 84 discriminative features. This SVM can completely separate the 47 short-term survivors and 26 long-term survivors; the lowest risk score assigned to the short-term survivors by the model is above 0.7, and most of the long-term survivors has a risk score lower than 0.3. In (Rosenwald *et al.*, 2002), the 80 validation samples are stratified according to the quartiles of the scores with each of quartiles consisting of 20 patients (p<0.001). To compare our results with those achieved in (Rosenwald *et al.*, 2002), we also partition patients into four risk groups but in a different way, defined as:

$$T\ is \begin{cases} \text{high risk}, & \text{if } S(T) > 0.7 \\ \text{intermediate-high risk}, & \text{if } 0.5 < S(T) \leq 0.7 \\ \text{intermediate-low risk}, & \text{if } 0.3 < S(T) \leq 0.5 \\ \text{low risk}, & \text{if } S(T) \leq 0.3 \end{cases} \quad (5)$$

where the threshold 0.5 is the mean value of 0.7 and 0.3. The overall survival Kaplan-Meier curves of the four risk groups are plotted in Figure 2 for the 80 validation samples.

We can see that the five-year survival rates for the high risk and low risk groups are clearly distinguishable. Though there is no distinct overall survival between the two intermediate groups, the 5-year survival rates of these two groups are obviously different from that in the high risk group or the low

Table 1. Comparison of the p-value (of log-rank test) obtained by applying different gene identification schemes and sample selection methods to the DLBCL data. "Selected SAM" means to use only top genes ranked by SAM score (cutoff value is 2.32 in this application). The number in the bracket under each gene identification scheme is the number of genes selected by the scheme.

With sample selection				
Validation data	All genes (7399)	Phase I (132)	Phase II (84)	Selected SAM (91)
80 samples	0.0125	0.0048	<0.0001	0.0015
167 samples	0.0069	0.0008	<0.0001	0.0005
Without sample selection				
Validation data	All genes (7399)	Phase I (88)	Phase II (30)	Selected SAM (100)
80 samples	0.2499	0.2788	0.7368	0.4513

risk group. This suggests that three groups would be appropriate for these DLBCL samples. So in the rest of this study, we merge intermediate-high and intermediate-low risk patients into a single intermediate risk category. Figure 3 shows Kaplan-Meier curves where we can see a significant survival difference for patients in each of our risk categories, for the 80 testing samples, for the 167 validation samples (80 testing samples plus 87 (=160-73) "non-extreme" samples in the original training set) and for the total 240 samples.

Other results on the validation samples obtained from this data set are reported in Table 1. These include p-values of the following tests: (i) using all features, features selected in Phase I, features selected by our two-phase method, or features selected by SAM based on the extreme patient samples; (ii) the above feature selection methods but based on the original training samples (taking status as class labels). From these results, we can see that using all training samples irrespective of the extreme cases can not achieve a good p-value no matter which feature selection method is applied. For more information, interested readers are referred to Figure F1 of our supplementary information to see Kaplan-Meier survival curves of these experiments (only for those with training sample selection). By the way, on the same data set, Bair and Tibshirani (2004) achieved p=0.00124 by categorising the patients into two risk groups using a semi-supervised machine learning approach.

Breast cancer

Currently, breast cancer patients with the same stage of disease have markedly different treatment responses and overall outcome (van't Veer *et al.*, 2002). The widely used clinical predictors for metastases, such as lymph node status and histological grade, can not provide accurate classifications for the tumors. Thus, more accurate methods of prognostication in breast cancer are needed to improve the selection of patients for adjuvant systemic therapy (van de Vijver *et al.*, 2002).

A comprehensive study for predicting the time to metastasis in breast cancer has been conducted by van't Veer *et al.* (2002) where a 70-gene prediction model has been developed using gene expression profile of 78 breast cancer patients. Those important genes were identified from more than 5000 genes using a complicated method. The method has the following steps: (i) first calculating the correlation coefficient of the expression for each gene with disease outcome and sorting the magnitude of the correlation coefficient to form a rank-ordered list, (ii) then sequentially adding subsets of 5 genes from the top of the list to the classification model, (iii) then evaluating the model using leave-one-out cross validation, (iv) repeating (ii) and (iii) until an optimal number of marker genes is reached. This 70-gene prediction model was re-used later in a separate study by van de Vijver *et al*. (2002) for analysing a bigger data set of 295 breast cancer patients.

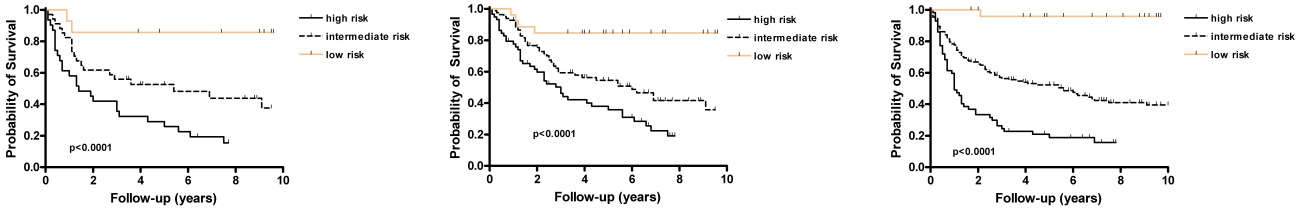
In our study, we use van de Vijver's data set. Note that this data set contains the 61 lymph-node-negative patients of van't Veer's data set. We conduct two kinds of analyses: metastasis and survival.

Metastasis prediction for breast cancer patients

Distant metastases are defined as a first event to be a treatment failure, and the data on patients is usually analysed from the date of surgery to the time of the first event (i.e. distant metastases or dead) or the date when the data is censored (van de Vijver *et al.*, 2002). Patients involved in metastasis study include those who had had distant metastases as a first event within five years and those who had remained free of disease for at least five years.

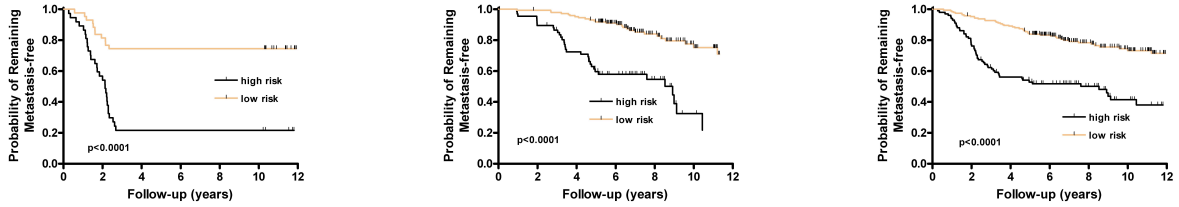
To select extreme cases, we set c_1=3 years and c_2=10 years in Formula (1). A total of 52 short-term survivors (i.e. who had distant metastases within three years) and 76 long-term survivors (i.e. who remained free of disease at least ten years) are among the 295 patients. As there is no independent validation data in this data set, we randomly selected 40 samples from each of these two types extreme cases to form our training set, and all the remaining samples (215 samples) are treated as validation data. We identified 9 genes from the 70 available genes based on the 80 training samples. The 9 genes are all selected by the entropy method in Phase I, and the Wilcoxon rank sum test does not remove any of them in Phase II. The constructed SVM model assigns a validation sample T to the high risk group if the risk score $S(T) > 0.5$, or otherwise to the low risk group if $S(T) \leq 0.5$. From the Kaplan-Meier curves drawn in Figure 4, we can see a significant difference in the probability that patients would remain free of distant metastases between the high and low risk groups of patients.

Results of using different gene identification schemes and our selected training samples are also good for this study — using all 70 genes or 31 SAM-selected genes can also achieve very small p-value (<0.0001) on the validation samples. Please refer to Figure F2 of our supplementary information to see Kaplan-Meier survival curves of these two tests. By the way, we find 6 common features selected by



(A) 80 testing samples (B) 167 validation samples (C) All 240 samples

Fig. 3. Kaplan-Meier plots illustrate the estimation of survival among different risk groups for a DLBCL study. (A) for the 80 testing samples, (B) for the 167 validation samples (80 testing samples plus 87 "non-extreme" samples in the original training set), and (C) for the entire 240 samples.



(A) Training samples (B) Validation samples (C) All samples

Fig. 4. Kaplan-Meier plots show the probability that patients would remain metastasis-free among different risk groups for a breast cancer study. (A) for the 80 training samples, (B) for the 215 validation samples, and (C) for all the 295 samples.

both our method and SAM. They are *Contig38288_RC, Contig55725_RC, NM_020974, NM_003981, NM_016359* and *X05610*.

Our result obtained on this data set is not directly comparable to that obtained by van de Vijver *et al*. (2002) because the 70-gene model they used was built on van't Veer's data. They have reported a good result of p<0.001 on these 295 samples. As mentioned, these 295 patients include 61 (of 78) training samples with lymph-node-negative. In a study on the same data, Bair and Tibshirani (2004) selected only five genes from 70 candidate using those 78 training samples. With a proposed supervised principal components method, they have achieved p<0.001 for all 295 patients and p=0.00328 for 234 patients excluding those used for training, respectively. The reason that we do not follow the same training and validating strategy is that we can not find clear indications in van de Vijver's data set for the 61 training samples.

Survival prediction for breast cancer patients

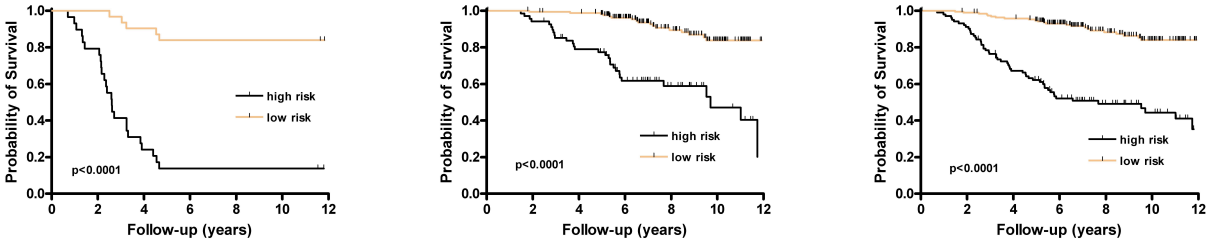
Besides the probability of remaining free of distant metastases, we also analyse the overall survival of breast cancer patients using gene expression profile of these 295 samples.

To select extreme cases, we set c_1=5 years and c_2=10 years in Formula (1). 48 short-term survivors and 83 long-term survivors are thus found among the 295 patients. Similar to the metastases study, we randomly selected 30 samples from each of these two types extreme cases to form our training set, all the remaining samples (235 samples) are treated as validation data. We identified 16 genes based on the 60 selected training samples. The constructed SVM model assigns a validation sample T to the high risk group if the risk score $S(T) > 0.5$, or otherwise to the low risk group if $S(T) \leq 0.5$. From the Kaplan-Meier curves drawn in Figure 5, we can see a significant difference in overall survival between the high and low risk groups of patients.

Results of using different gene identification schemes and our selected training samples are also good for this study — using all 70 genes or 35 SAM-selected genes can also achieve very small p-value (<0.0001) on the validation samples. Please refer to Figure F3 of our supplementary information to see Kaplan-Meier survival curves of these two tests. In addition, we find 14 common genes selected by both our method and SAM. They are *NM_007203, NM_005915, Contig38288_RC, Contig55725_RC, Contig46223_RC, NM_020974, NM_016577, Contig35251_RC, NM_014791, NM_003981, NM_006681, X05610, NM_000849* and *Contig56457_RC*.

Adult acute myeloid leukemia

Currently, the prognostic indicators to identify the appropriate therapy for acute myeloid leukemia (AML) patients include age, cytogenetic findings, the white-cell count and the presence or absence of recurrent cytogenetic aberrations (Bullinger *et al*., 2004). However, these factors do not fully reflect the molecular heterogeneity of the disease and treatment stratification is difficult. Thus, predictors built on gene



(A) Training samples (B) Validation samples (C) All samples

Fig. 5. Kaplan-Meier plots show the overall survival among different risk groups for a breast cancer study. (A) for the 60 training samples, (B) for the 235 validation samples, and (C) for all the 295 samples.

expression profiles are expected to accurately predict the clinical outcome at molecular level so that appropriate treatment can be tailored for individual patient.

Bullinger *et al.* (2004) have studied gene expression in peripheral-blood samples or bone marrow samples from 116 adults with AML and identified new molecular subtypes of AML by unsupervised hierarchical clustering analysis. They have randomly divided these 116 samples into a training set containing 59 samples and a test set containing 57 samples — with a similar number of samples in each set are from patients who had died. In the training set, 26 patients were alive with follow-up time from 138 days to 1625 days, and 33 were dead with follow-up time from 1 day to 730 days.

To select extreme cases, we set c_1=1 year and c_2=2 years in Formula (1). A total of 29 short-term survivors and 8 long-term survivors are found among the 59 training samples. So, our training set consists of only these 37 samples. From these 37 extreme patient samples, we identified 50 features that are related to patient survival status by using our feature filtering method. The constructed SVM model assigns a validation sample T to the high risk group if the risk score $S(T) > 0.5$, or otherwise to the low risk group if $S(T) \leq 0.5$. The Kaplan-Meier curves in Figure 6 shows a significant difference in overall survival between the high and low risk groups of patients: for the 57 testing samples, for the 79 validation samples (including 22 "non-extreme" cases in the original training set), and for the entire 116 samples.

For this data set, we also obtained the results of using different feature selection schemes or without training sample selection. In Table 2, p-value for each of these tests are listed. Kaplan-Meier survival curves for some of these experiments can be found in Figure F4 of our supplementary information. Generally, we use similar number of SAM-selected genes as that selected by our method.

On the same data set, Bullinger *et al.* has applied 149 SAM-selected cDNAs that identified from all training samples and a clustering method to estimate the outcome. They has reported a good result (p=0.006) on overall survival of the patients in their *poor-outcome* and *good-outcome* groups. We tried to feed same number of SAM-selected features to our model, but we only obtained a result of p=0.0133 by using all training samples (see results in Table 2). However, we achieved better performance by using our selected training samples. In addition, our results are also better than that (p=0.00136) reported in (Bair and Tibshirani, 2004) on the same data set.

Table 2. Comparison of the p-value (of log-rank test) obtained by applying different gene identification schemes and sample selection methods to the AML data. "Selected SAM" means to use number of top genes ranked by SAM score. The number in the bracket under each gene identification scheme is the number of genes selected by the scheme.

With sample selection					
Validation data	All genes (6283)	Phase I (61)	Phase II (50)	Selected SAM (100)	Selected SAM (50)
57 samples	0.0133	0.0008	0.0007	0.0015	0.0015
79 samples	0.0230	0.0020	0.0024	0.0020	0.0045
Without sample selection					
Validation data	All genes (6283)	Phase I (133)	Phase II (80)	Selected SAM (149)	Selected SAM (80)
57 samples	0.2938	0.1843	0.0889	0.0133	0.1478

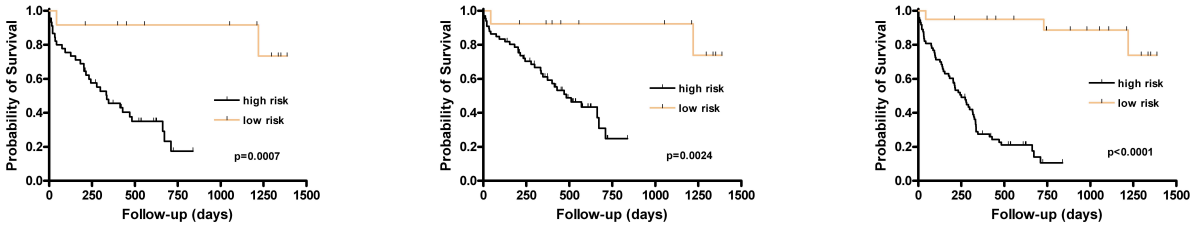
Table 3. Number of samples in original training data and selected training set of the DLBCL and AML data sets.

Application	Original training set			Our training set		
	Alive	Dead	Total	Long-term	Short-term	Total
DLBCL	72	88	160	26	47	73
AML	26	33	59	8	29	37

DISCUSSION

Recall that we select only long-term and short-term survivors for training the prediction models. Table 3 lists the size change from the original training samples to our selected training set for the DLBCL and AML applications. Oberve that our method uses roughly half of the samples as training. As already shown in Table 1 and Table 2, these informative training samples can indeed make performance improvement, even using SAM for gene selection.

As discussed in Section METHOD, we have some basic guide lines to determine the thresholds c_1 and c_2 that defined in Formula (1). Bearing these minimum constraints in



(A) 57 testing samples (B) 79 validation samples (C) All 116 samples

Fig. 6. Kaplan-Meier plots estimate the overall survival among different risk groups for an adult acute myeloid leukemia study. (A) for the 57 testing samples, (B) for the 79 validation samples (57 testing samples plus 22 "non-extreme" cases in the original training set), and (C) for the entire 116 samples.

mind, we have tried several different c_1 and c_2 values in our study. In Table 4, p-value (of the log-rank test) associated with the Kaplan-Meier survival curves on validation samples under different selections of the c_1 and c_2 from DLBCL study are listed. All results are based on the selected genes using our gene identification scheme. We can see that for a range of c_1 and c_2 (i.e. $c_1 < 3$ years and $c_2 \geq 7$ years), we can achieve better predictions by selecting extreme samples. In any case, the selection of c_1 and c_2 can be further refined by running cross-validation on training samples.

To demonstrate the effectiveness of selecting extreme samples, we have also done following tests. (i) Using only "non-extreme" samples in the original training set to build prediction model. As expected, the results are not good. For example, in DLBCL study, there are 87 "non-extreme" samples left after we select 73 extreme cases from 160 samples in the preliminary group of the data. When we use these samples to train model, we get p=0.4481 (40 features selected by our method) and p=0.5887 (all genes) on 80 validation cases. (ii) Incorporating the idea of transductive SVM (tSVM) to include those "non-extreme" cases into training data as unlabeled samples to build prediction model. The results are not better than those we presented above. In the AML study, tSVM achieves p=0.0574 (50 features selected by our method), p=0.0487 (top 100 features selected by SAM) and p=0.0468 (all genes) on the 57 validation samples. In the DLBCL study, tSVM achieves p=0.0044 (84 features selected by out method), and p=0.0113 (all genes) on the 80 validation samples. The software we used is SVM^{light} (version 6.01, `http://svmlight.joachims.org/`).

According to our experience on gene expression data analysis, the entropy measure can filter out about 90-95% of total number of genes (Li *et al.*, 2003). This point has been verified again in our outcome prediction: for example, the entropy measure retains only 61 features (out of total 6283 candidates) in AML study. After further being filtered by Wilcoxon rank sum test, only 50 of them are kept to build prediction model. Most importantly, these selected genes achieve better experimental performance — using only Wilcoxon rank sum test to select features (i.e. no Phase I) will lead to a larger number of selected genes but worse results. For example, if we only apply rank sum test to the 70 genes provided by the breast cancer data, 40 genes will be selected for metastasis prediction and 42 genes selected for survival prediction. With these larger number of genes, the p values for testing on validation samples are only 0.0004 (metastasis) and 0.0007 (survival), respectively. This observation indicates that only applying rank sum test may not be powerful enough to reduce number of genes and thus, we use it at second phase after more than 90% genes have been removed by the entropy-based algorithm. Table 5 shows in DLBCL and AML studies, the number-change trend of features from original data, to the entropy selection (Phase I) and to Wilcoxon rank sum test (Phase II), as well as to applying only the rank sum test. It can be seen that the feature reduction is mostly by the entropy selection.

In the current study, a simple linear kernel SVM is trained on the selected samples and genes to build a scoring model.

Table 4. Results of using different thresholds c_1 (years) and c_2 (years) in training sample selection on DLBCL study. All results are based on our proposed gene identification scheme and on validation samples only. Column "Short-term"/"Long-term" means the number of short-term/long-term survivors.

c_1	c_2	p-value	Short-term	Long-term	No. genes
1	5	0.2962	47	57	121
1	7	0.0110	47	36	79
1	8	<0.0001	47	26	84
1	9	0.0570	47	22	40
2	8	0.0047	61	26	55
3	8	0.0761	76	26	51

Table 5. Number of genes left after feature filtering for each phase of our gene identification scheme and for only applying Wilcoxon rank sum test (i.e. RSTOnly) in the DLBCL and AML studies. The percentage in the brackets indicates the proportion of the remaining genes on original feature space.

Data set	Orignal	Phase I	Phase II	RSTOnly
DLBCL	7399	132 (1.8%)	84 (1.1%)	3632 (49.1%)
AML	6283	61 (1.0%)	50 (0.8%)	252 (4.0%)

The model then assigns each validation sample a risk score to predict patient outcome. Based on the training results, we can derive explicit thresholds (e.g., 0.5, 0.3, 0.7) of our risk score to categorize patients into different risk groups. Thus, when a new case comes, we are able to assign it to the corresponding risk group easily according to its risk score.

In summary, we have applied statistical and machine learning technologies to predict patient outcome from gene expression profiles and clinical information. Different from other works, we pick out extreme cases to form the training set, consisting of only short-term survivors who died within a short period and long-term survivors who were still alive after a relevant long follow-up time. Our in-silico experimental results on three public gene expression data sets have demonstrated the high effectiveness of our idea.

We have some ongoing studies. (1) Data sets from other tumors are under analysis. (2) In order to obtain a more refined set of genes, some matrices to measure the correlation between the genes selected by our two-phase filtering method are under testing — correlation test will be conducted within each of the two groups of genes left by the Wilcoxon rank sum test (i.e. one group contains genes whose statistical measure less than the lower bound critical value and one group contains genes whose statistical measure larger than the upper bound critical value). (3) Directly predict patient survival time using regression algorithms. But one concern is that those alive patients with short-term follow-ups may not be useful for this direct regression approach.

ACKNOWLEDGEMENT

We would like to acknowledge the two anonymous reviewers for their many valuable comments on the manuscript.

REFERENCES

Altman, D.B. (1991) *Practical statistics for medical research*, Chapman and Hall.

Ando, T., Katayama, M. (2002) Selection of causal gene sets from transcriptional profiling by FNN modeling and prediction of lymphoma outcome. In *13th Intl. Conf. Genome Informatics*, pp. 278-279.

Bair, E., Tibshirani, R. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, **2**(4):0511-0522.

Beer, D.G., Kardia, S.L., Huang, C.C. *et al*. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**(8):816-823.

Bullinger, L., Dohner. K., Bair, E. *et al*. (2004) Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *NEJM*, **350**(16):1605-1616.

Chu, G., Narasimhan, B., Tibshirani, R. and Tusher, V.G. (2004) *SAM user guide and technical document*., (`http://www-stat.stanford.edu/~tibs/SAM/.`)

Cox, D.R. (1972) Regression models and life-tables (with discussion). *J. R. Stat. Soc*., **B34**:187-220.

Fayyad, U., Irani, K. (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In *13th Intl. Joint Conf. Artificial Intelligence*, 1022-1029.

LeBlanc, M., Kooperberg, C., Grogan, T.M., Miller, T.P. (2003) Directed indices for exploring gene expression data. *Bioinformatics*, **19**(6):686–693.

Li, J., Liu, H., Wong, L. (2003) Mean-entropy discretized features are effective for classifying high-dimensional biomedical data. *3rd ACM SIGKDD Workshop on Data Mining*, 17-24.

Liu, H. & Setiono, R. (1995) Chi2: Feature selection and discretization of numeric attributes. In *Proc. of 7th IEEE Intl. Conf. on Tools with Artificial Intelligence*, 338-391.

Lunn, M., McNeil, D.R. (1995) Applying Cox Regression to Competing Risks. *Biometrics*, **51**:524-532.

Park, P.J., Pagano, M., Bonetti, M. (2001) A non-parametric scoring algorithm for identifying informative genes from microarray data. *Pac. Symp. Biocomput.*, 52-63

Park, P.J., Tian, L., Kohane, S. (2002) Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18**(Suppl 1):S120-S127.

Rosenwald, A., Wright, G., Chan, W.C. *et al*. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma, *NEJM*, **346**(25):1937-1947.

Shipp, M.A., Ross, K.N., Tamayo, P. *et al*. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning *Nat. Med.*, **8**(1):68-74.

Troyanskaya, O.G., Garber, M.E., Brown, P.O. *et al*. (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**(11): 1454-1461, 2002.

Tusher, V.G., Tibshirani, R., Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, **98**:5116-5121.

Vapnik, V.N. (1995) *The Natural of Statistical Learning Theory*, Springer.

van de Vijver, M.J., He, Y.D., van't Veer, L.J. *et al*. (2002) A gene-expression signature as a predictor of survival in breast cancer. *NEJM*, **347**(25):1999-2009.

van't Veer, L.J., Dai, H., van de Vijver, M.J. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871):530-536.

Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods. *Biometrics*, **1**, 80-83.

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A. *et al*. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**:133-143.