

# Selection of Patient Samples and Genes for Outcome Prediction

Huiqing Liu Jinyan Li Limsoon Wong  
Institute for Infocomm Research  
21 Heng Mui Keng Terrace  
Singapore 119613  
{huiqing, jinyan, limsoon}@i2r.a-star.edu.sg

## Abstract

*Gene expression profiles with clinical outcome data enable monitoring of disease progression and prediction of patient survival at the molecular level. We present a new computational method for outcome prediction. Our idea is to use an informative subset of original training samples. This subset consists of only short-term survivors who died within a short period and long-term survivors who were still alive after a long follow-up time. These extreme training samples yield a clear platform to identify genes whose expression is related to survival. To find relevant genes, we combine two feature selection methods — entropy measure and Wilcoxon rank sum test — so that a set of sharp discriminating features are identified. The selected training samples and genes are then integrated by a support vector machine to build a prediction model, by which each validation sample is assigned a survival/relapse risk score for drawing Kaplan-Meier survival curves. We apply this method to two data sets: diffuse large-B-cell lymphoma (DLBCL) and primary lung adenocarcinoma. In both cases, patients in high and low risk groups stratified by our risk scores are clearly distinguishable. We also compare our risk scores to some clinical factors, such as International Prognostic Index score for DLBCL analysis and tumor stage information for lung adenocarcinoma. Our results indicate that gene expression profiles combined with carefully chosen learning algorithms can predict patient survival for certain diseases.*

## 1. Introduction

Microarray technology permits monitoring of the expression levels of thousands of genes simultaneously. A few previous studies have shown promising results for outcome prediction using gene expression profiles for certain diseases [11, 3, 16, 20, 10, 7]. This kind of analysis provides techniques to predict disease progression and clinical outcome at the molecular level. It also identifies genes whose

expression is most related to survival. Carefully verifying and understanding these active genes would lead to innovative therapies and may also generate opportunities for drug discovery.

Various approaches have recently been used on outcome prediction using gene expression profiles. In the Cox proportional hazard regression method [5, 9], genes most related to survival are first 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 [3, 11]. In Ando *et al* [2], gene expression profiles are fed to a fuzzy neural network (FNN) system to predict survival of patients. They first predict the outcome of each patient using one gene at one time. Then they rank each gene by their accuracy. Next, one by one, they use the ten highest ranked genes and the selected partner genes for prediction. Finally, the formed ten FNN models using combinatorial genes are optimized by the back-propagation method. In Park *et al* [10], gene expression data are linked to patient survival times using the partial least squares regression technique, which is a compromise between principal component analysis and ordinary least squares regression. In Shipp *et al* [14], the weighted voting algorithm is used to identify cured *versus* fatal for outcome of diffuse large B-cell lymphoma. The algorithm calculates the weighted combination of selected informative marker genes to make a class distinction. In [7], LeBlanc *et al* develop a gene index method to investigate genes that jointly relate to patient outcome and to a specific “reference gene” of interest.

We present here a new computational method for outcome prediction using gene expression profiles. Different from all previous works, in the first step, we carefully form the training set samples by selecting 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. This idea is motivated by our belief that short-term and long-term survivors are more informative and reliable (than

those cases in between) for building and understanding the relationship between genes and patient outcome. In the second step, to identify genes most associated with outcome, a two-phase feature selection procedure — combining entropy measure and Wilcoxon rank sum test — is applied to the training set. After this filtering, the number of resulting genes are expected to be the reasonable representatives to link gene expression profiles to patient outcome. In the third step, a linear kernel support vector machine (SVM) is trained on the selected samples and genes to build a scoring model. The model assigns each validation sample a risk score to predict patient outcome. Explicit threshold values to categorize different risk groups are easily obtained based on training results, so that outcome prediction for the new coming cases is possible.

We test the effectiveness of our method on two applications: predicting survival of patients after chemotherapy for diffuse large-B-cell lymphoma (DLBCL), and predicting outcome of primary lung adenocarcinoma. The corresponding Kaplan-Meier survival curves illustrate that patients assigned to the different risk groups based on our score have significant difference on survival. We also analyse the relationship between our risk group and some clinical factors, such as International Prognostic Index score for DLBCL analysis, and tumor stage information for lung adenocarcinoma. Our good results indicate that gene expression profiles and supervised machine learning techniques can be used to predict patient outcome for certain diseases.

## 2. Methods

In this section, we begin with a new idea to select an informative subset of training samples, then we describe how to identify relevant genes based on these training samples, then we propose a scoring function for survival risk estimation and survival prediction.

### 2.1. Selection of informative training samples

Let  $D$  be a training data set for a survival study, then  $D$  usually contains two classes of samples  $D_{died}$  and  $D_{alive}$ . Here  $D_{died}$  is the set of patients who died within  $x$  years, and  $D_{alive}$  is the set of patients who were still alive after  $x$  years. A widely used value for  $x$  is 5 (years). An informative subset of training samples for  $D$  is then the union of a subset of  $D_{died}$  and a subset of  $D_{alive}$ . The subset of  $D_{died}$  are those patients who died in a *short* period (e.g., 1 year) — named as “*short-term survivors*”; while the subset of  $D_{alive}$  are those patients who were alive after a *long* period (e.g., 8 years) — named as “*long-term survivors*”. Note that long-term survivors may include those patients who died after the specified long period. The short-term and long-term survivors are called extreme cases.

This idea emphasizes that the extreme cases play more important roles for survival predication than those in the “middle” status. I.e., we do not expect reliable prediction could come out from analysing alive patients whose available follow-up time is short. This idea also helps the identification of those genes that are closely relevant to survival.

Formally, 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

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

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. The two thresholds can vary from disease to disease, from data set to data set. For example, in the survival study of early-stage lung adenocarcinoma, we choose short-term survivors as those who died within one follow-up year (i.e.  $c_1$  is 1 year) and long-term survivors as those who are alive after five follow-up years (i.e.  $c_2$  is 5 years). There are total 31 extreme training samples (10 short-term survivors and 21 long-term survivors) among total of 86 available primary lung adenocarcinoma patients. These 21 long-term survivors include 2 patients whose status at the end of follow-up time was “dead”, but follow-up time was 79.5 months and 84.1 months, respectively. Our basic guideline for  $c_1$  and  $c_2$  is that the informative subset is between one third and one half of total available samples.

### 2.2. Identification of relevant genes

In this section, we describe our two-phase feature selection procedure to identify genes expressed differentially between short-term and long-term survivors, where we combine two basic feature selection methods: entropy measure [6] and Wilcoxon rank sum test [17]. The purpose is to identify a subset of sharp discriminating features. The entropy measure is effective for identifying discriminating features. After narrowing down by the Wilcoxon rank sum test, the remaining features become sharply discriminating. This systematic method usually selects less than 10% of the original features.

In phase I, we apply Fayyad’s discretization algorithm [6] to all the genes. The algorithm partitions the value range of a numeric feature such that each of the resulting intervals contain the same class of samples, as many as possible. See Appendix A for details of the algorithm. If the algorithm cannot find a suitable cut point

to split a feature's value range, then this feature is removed from our consideration.

In phase II, we conduct the Wilcoxon rank sum test on genes kept by phase I. For each gene  $X$ , a test statistical measure  $w(X)$  is calculated via the manner described in Appendix B. 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 given in equation (12) in Appendix B, then we remove  $X$  from further consideration. Otherwise, we keep feature  $X$  because this feature 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  $c_{upper}$ , 5% or 1% significant level is generally used. We use 5% in this paper. See Figure 1 for a whole picture of gene identification and selection by our two-phase feature filtering procedure.

### 2.3. Construction of an SVM scoring function

In this section, we propose a scoring function to estimate the survival risk for every patient. This regression scoring function is based on support vector machines (SVM) [15]. In this study, we use the implementation of SVM in *Weka* version 3.2 (<http://www.cs.waikato.ac.nz/ml/weka>). We choose simple linear kernel function under which the final SVM regression function  $G(T)$  is a linear combination of the expression values of the identified genes, namely,

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

where the vectors  $x(i)$  are the support vectors,  $y_i$  are the class labels (1 and -1 used here) of  $x(i)$ , vector  $T$  represents a test sample, and  $\alpha_i$  and  $\beta$  are numeric parameters to be learned from the training data.

We map class label of "short-term survivors" to 1 and "long-term survivors" to -1. Note that  $G(T) > 0$  if the sample  $T$  is more likely to be a "short-term survivor", and  $G(T) < 0$  if the sample  $T$  is more likely to be a "long-term survivor". To normalize  $G(T)$ , we use a transformation function  $S(T)$  defined as:

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

So,  $G(T)$  is normalized by  $S(T)$  into the range (0, 1). Note that the smaller the  $S(T)$  value is, the better survival the corresponding patient  $T$  will have. We term  $S(T)$  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  $S(T)$ , where if  $S(T) > 0.5$  then the patient corresponding to sample  $T$  will have high risk; otherwise, the patient will have low risk. If more than two risk groups are considered — such as high,

intermediate, and low — then other cutoffs can be set based on the risk scores of 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,

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

Generally,  $r_1 > 0.5$ ,  $r_2 < 0.5$ , and they can be derived from the risk scores assigned to the training samples.

After assigning patients into different risk groups, we draw Kaplan-Meier plots [1] to compare the survival characteristics between groups. For each plot, the associated  $p$ -value is for the null hypothesis that the survival curves are no difference between two groups. Here, all the Kaplan-Meier survival curves are generated by *GraphPad Prism* (<http://www.graphpad.com>).

## 3. Results

In this section, we apply our method to two data sets and report the results.

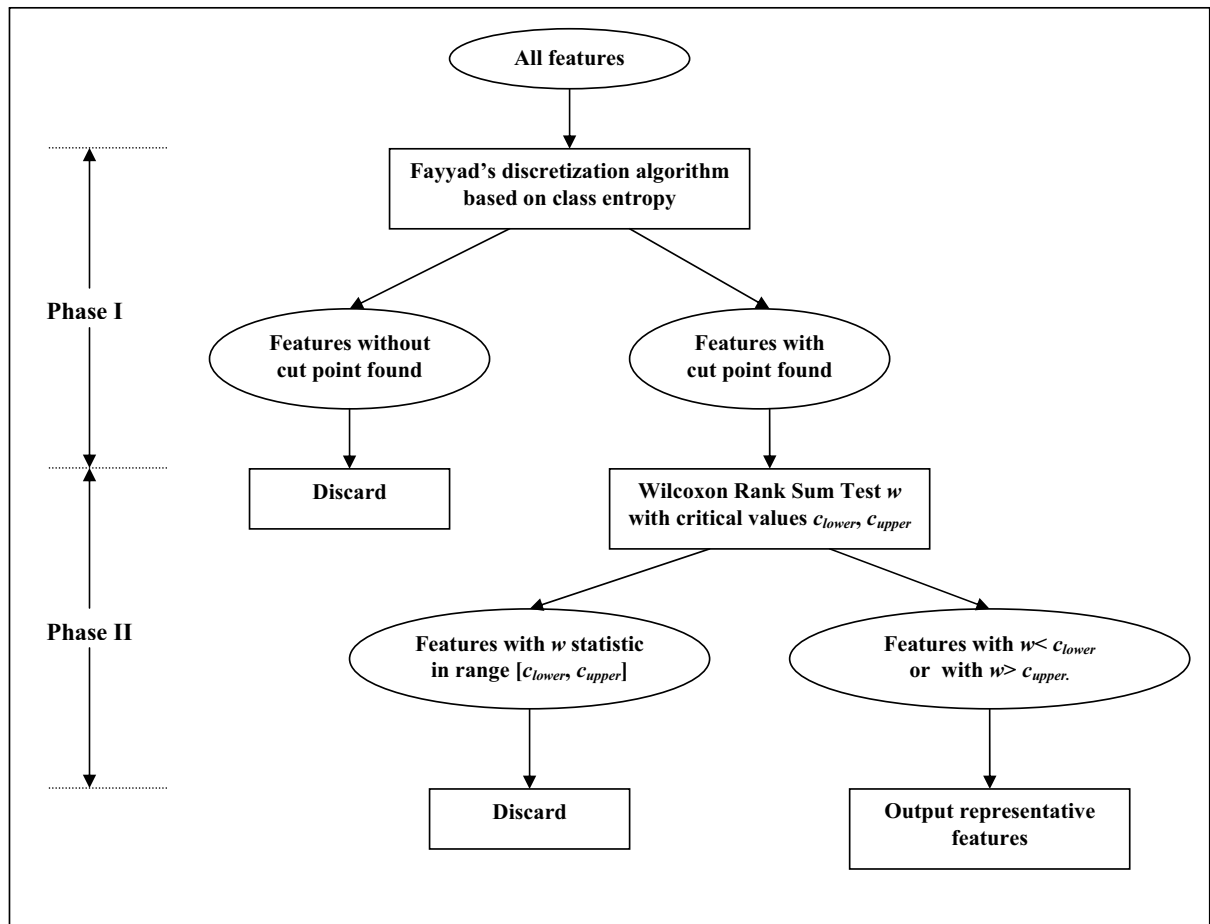
### 3.1. Lymphoma

Survival after chemotherapy for diffuse large-B-cell lymphoma (DLBCL) patients was previously studied by [11] using gene expression profiling and Cox proportional-hazards model. In that study, expression profiles of biopsy samples from 240 patients were used [11]. 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.

**Survival curves showing clear distinction.** As an initial step, we pre-process the data to remove those genes that are absent in more than 10% of the experiments in the preliminary group. This yields 4937 features.

Then, we select short-term survivors and long-term survivors to construct an informative subset of training samples. For this study, we set  $c_1 = 1$  year and  $c_2 = 8$  years in Formula (1). Among the preliminary 160-patient group, 47 short-term survivors (who died within one follow-up year) and 26 long-term survivors (who were alive after eight follow-up years) are thus chosen. So, a total of 73 samples are in this informative subset of training samples.

In the second step, we apply feature selection to these 73 samples and identify 84 genes that are related to patient survival status at 5% significant level (for Wilcoxon rank sum test). Some of our selected genes are also listed in Table 2 of [11], where these genes are found to be significantly associated with survival ( $p < 0.01$ ). E.g., AA805575 (GenBank accession number) is in *germinal-center B-cell signature*, X00452 and M20430 in *MHC class II signature*,



**Figure 1. A diagram of our two-phase feature filtering procedure — combining concepts of entropy measure (Phase I) and Wilcoxon rank sum test (Phase II).**

and D87071 is in *lymph-node signature*. The gene signatures were formed by a hierarchical clustering algorithm in [11]. Besides, some top-ranked genes (with smaller entropy value) identified by us are also in one of these gene signatures. E.g., 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*.

In the third step, an SVM model is trained on the 73 extreme training samples with the 84 identified features. We find that the well-learned linear kernel SVM can separate the 47 short-term survivors and 26 long-term survivors completely — the lowest risk score assigned to the short-term survivors is above 0.7 and most of the long-term survivors has risk score lower than 0.3. Then, we calculate risk scores  $S(T)$  for all the other samples, namely the remaining (non-extreme) 87 samples in the original preliminary group and

the 80 samples in the validation group. These 167 samples are treated as our test set.

We categorized patients into four risk groups as follows:

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

where the threshold 0.5 is the mean value of 0.7 and 0.3. The Kaplan-Meier curves of overall survival are drawn in Figure 2, where we can see clear differences at the five-year survival rates for the high risk and low risk groups, in both testing sample set (Panel (A)) and all samples (Panel (B)). Although we cannot see 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 risk group. This also suggests that three or two risk groups would be sufficient for these

DLBCL samples. So in the rest of this study, we simply merge high and intermediate-high risk patients into a single high risk category, and low and intermediate-low risk patients into a single low risk category.

Having the risk score, when a new case comes, we will be able to assign it to the corresponding risk group easily. This kind of prediction was not addressed in [11] where the DLBCL patients were ranked by their gene-expression-based outcome-predictor score but divided into several groups with equal number of samples. For an example: 80 samples in the validation group were stratified according to the quartiles of the scores with each of quartiles consisting of 20 patients. With this kind of categorization, one cannot find an explicit measure to evaluate a new case.

**Comparison with International Prognostic Index.** Various clinical features — such as stage, performance status, lactate dehydrogenase levels — which are known to be strongly related to patient survival, have been combined to form the International Prognostic Index (IPI) [13]. The IPI has been effectively adopted to separate aggressive lymphomas into several groups with significantly different responses to therapy and survival. Since IPI is only built on the consideration of clinical factors, it provides little insight into disease biology [7].

The risk score obtained from our method is based on gene expression in biopsy specimens of the lymphoma, so it is an independent predictor from IPI. In fact, we find that patients in the high IPI group — and similarly for the intermediate and the low IPI groups — when partitioned by our risk score into high risk and low risk categories, have significantly different outcomes. In Figure 3, Kaplan-Meier plots show significant difference on overall survival for our high risk and low risk groups among the patients with IPI high (and similarly for intermediate) risk index. In particular, among 21 IPI high risk patients in our testing set, 15 of them are assigned by our method to the high risk category and 6 of them to the low risk category. When we check to the survival status of these patients, we find 14 of the 15 patients belonging to our high risk category are indeed dead while only 2 of the 6 patients belonging to our low risk category are dead. Similarly, for all 32 patients in the whole data set with high IPI, 23 of them (22 dead) are assigned by our method to the high risk category and 9 (5 dead) of them are assigned to low risk category. This suggests that our method may be a more effective predictor of DLBCL survival outcome than the IPI.

### 3.2. Lung adenocarcinoma

Most patients with non-small cell lung cancer (NSCLC) present with advanced disease and the overall 10-year survival rate remains at 8-10% [3]. Adenocarcinoma is the

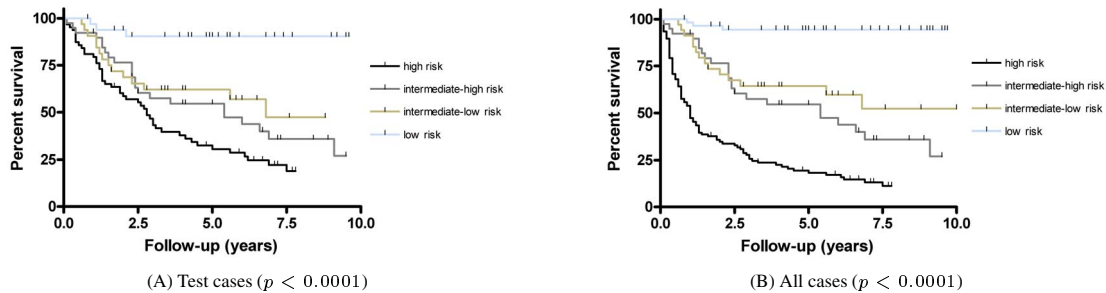
major histological subtype of non-small cell lung cancer (NSCLC). There is a need to better predict tumor progression and clinical outcome in lung adenocarcinoma. The lung adenocarcinoma data set contains 86 primary lung adenocarcinoma. These experiments include 67 stage I and 19 stage III tumors, each of them is described by 7129 genes. The data set was first analysed in [3] where a risk index was derived based on the top 50 good genes that were identified to be most related to survival by univariate Cox analysis. In that study, tests were conducted by randomly splitting 86 samples into equal sized training and testing sets and “leave-one-out” cross validation.

As per Section 2, we form our training set by setting  $c_1 = 1$  year and  $c_2 = 5$  years in Formula (1). 10 short-term survivors and 21 long-term survivors are thus chosen. Applying feature selection to these 31 training samples, we find 591 genes that are related to outcome. Our top-ranked feature by entropy measure, the ATRX gene, is a putative transcription regulator. It is also reported by Borczuk *et al* in their recent paper [4] on NSCLC. Our second-ranked feature, ENPP2, is part of stress pathways involved in oogenesis. Yang *et al*[19] also detected it in NSCLC.

Then we train a linear kernel SVM to obtain the weight for each identified gene based on training data. The trained SVM can separate these 31 samples very well, assigning very high risk scores to short-term survivors (lowest score is as high as 0.73) while very low risk scores to long-term survivors (highest score is as low as 0.25).

After training, we calculate risk score for each of the remaining 55 samples which are used for test purpose. These samples are then classified as high risk group consisting samples  $T$  with  $S(T) > 0.5$ , or as low risk group consisting samples  $T$  with  $S(T) \leq 0.5$ . The Kaplan-Meier curves in Figure 4 show clear difference of survival for patients in our high and low risk groups for both testing cases and all cases. Since we pick out all short-term and long-term survivors to form the training set, there is no “death” event happened in the first 12 months time and no sample censored after 60 months time in the plot drawn only on the test cases (Panel (A)).

In order to understand the relationship between our prediction and tumor stage (I or III). We also draw Kaplan-Meier curves to delineate survival difference between our high and low risk patients conditioned on tumor stage. From Figure 5, we can see that outcomes of patients with stage I lung adenocarcinoma in our high and low risk groups differ from each other, for both test cases (Panel(A)) and all cases (Panel(B)). Again remarkably, for 13 stage III cases in the testing set, we assigned 11 of them to high risk group, and the 2 of them assigned to low risk group were all alive at the end of the follow-up time. Among all 19 stage III cases, 17 of them were assigned to high risk group according to our risk score.



**Figure 2. Kaplan-Meier plots illustrate the estimation of overall survival among different risk DLBCL patients in the testing set containing 167 samples (Panel (A)) and all 240 samples (Panel (B)). The risk groups are formed on our SVM-based scoring function. A tick mark on the plot indicates that one sample is censored at the corresponding time. The 5-year overall survival for high risk versus low risk groups of patients for testing samples is 32% versus 91%, for all samples is 20% versus 95%.**

#### 4. Discussion and Summary

In the step of training set construction, we select only two extreme cases — long-term and short-term survivors. See Table 1 for size change trends from the original training samples to the informative training samples on DLBCL and lung adenocarcinoma data sets. The figures illustrate that we used a small part of samples as training.

| Application         | Data set    | Status   |       | Total |
|---------------------|-------------|----------|-------|-------|
|                     |             | Dead     | Alive |       |
| DLBCL               | Original    | 88       | 72    | 160   |
|                     | Informative | 47+1(*)  | 25    | 73    |
| Lung adenocarcinoma | Original    | 24       | 62    | 86    |
|                     | Informative | 10+2(**) | 19    | 31    |

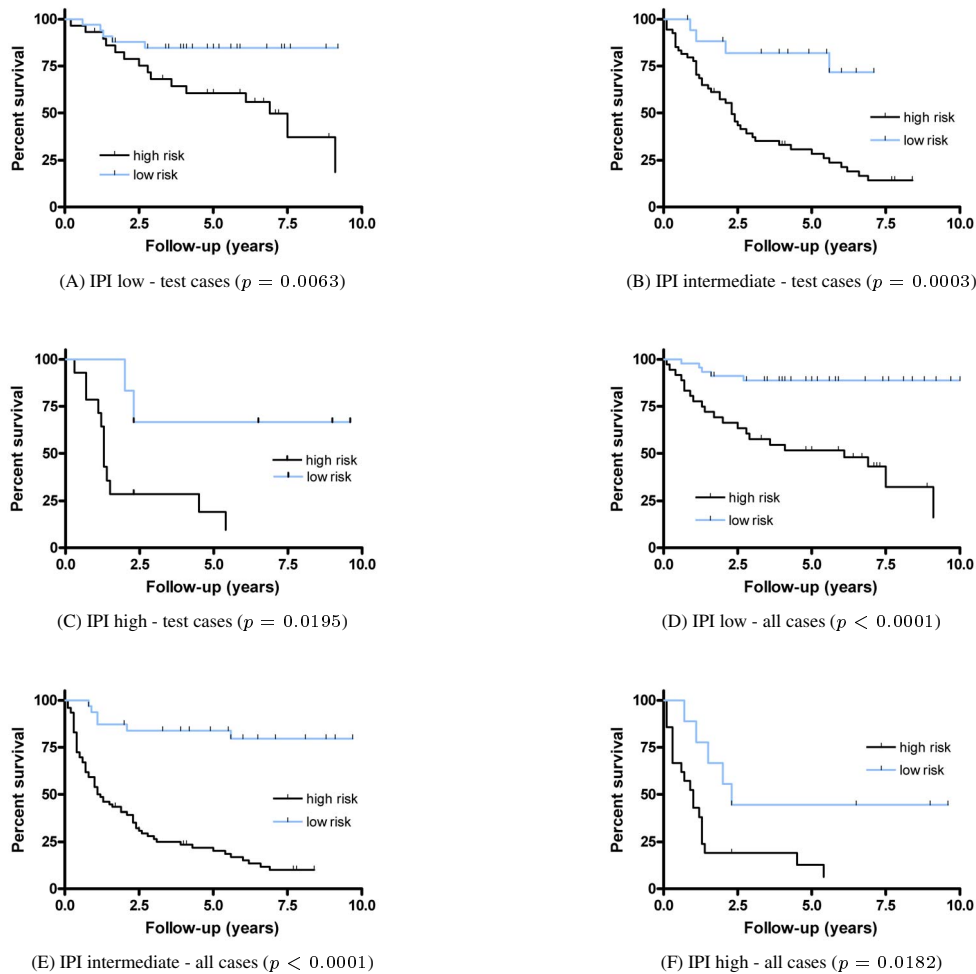
**Table 1. Number of samples in original data and selected informative training set. (\*) There are 48 DLBCL samples, whose relevant patient was dead at the end of follow-up time, are selected as informative, 47 of them are short-term survivors while 1 of them is long-term survivor. (\*\*) There are 12 lung adenocarcinoma, whose relevant patient was dead, are selected as informative, 10 of them are short-term survivors while 2 of them are long-term survivors.**

On the other hand, if we do not select those extreme cases, and instead use all available training samples, then what will be the results? To illustrate this, we select genes and train SVM model on the 160 samples in the prelimi-

nary group of DLBCL study. Although the training accuracy is good, Kaplan-Meier plots do not show significant survival difference between the high and low risk groups formed by the 80 validation samples based on their risk scores that assigned by the trained SVM model. In detail, using all 4937 genes, the  $p$  value of the survival curves is 0.21 ((A) in Figure 6); using 88 genes selected by our feature filtering method, the  $p$  value is 0.38 ((B) in Figure 6). Therefore, we claim that our proposed idea of selecting informative training samples is an effective method.

In the step of gene identification, built on statistical knowledge, our two-phase filtering process discards many unrelated genes and only keeps a small number of informative representatives. According to our experience on gene expression data analysis, generally, entropy measure can filter out about 90% of total number of genes [8]. This point has been verified again in outcome prediction: entropy measure retains only 132 genes in DLBCL study (there are around 5000 genes after removing missing values) and 884 genes in lung adenocarcinoma study (original data contain 7129 genes). In fact, these genes can also lead to good results. After further filtering by Wilcoxon rank sum test, the final selected genes are with smaller size. Most importantly, these genes achieve better experimental performance. Table 2 shows the number-change trend of features from original data to the entropy selection, and to Wilcoxon rank sum test selection. It can be seen that the feature reduction is mostly by the entropy selection.

For comparison, in DLBCL study, we also do experiments using all the 4937 genes and the 132 genes output from the Phase I filtering. The results show that in each of these cases, the overall survival difference between the high and low risk groups formed by our risk scores on the testing samples can be seen as well. In Figure 7, we draw the cor-



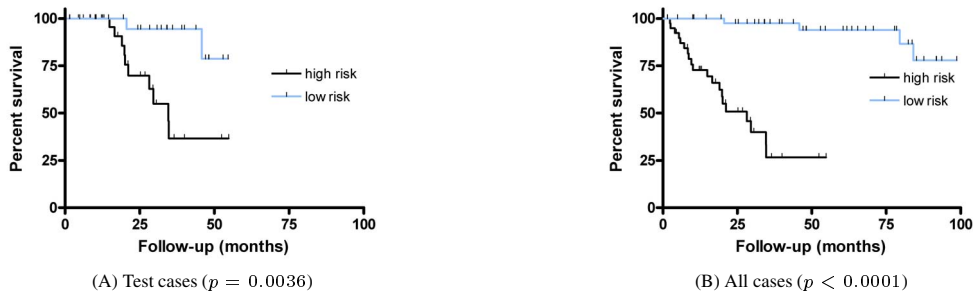
**Figure 3. Kaplan-Meier Estimates of survival among high risk and low risk DLBCL patients (according to our method) in each IPI defined group. Plots (A), (B) and (C) are based on 167 testing samples while (D), (E) and (F) are for all 240 cases.**

responding Kaplan-Meier survival curves. Again, the good results also demonstrate the effectiveness of selection the informative samples. In addition, in the study of lung adenocarcinoma, using all genes (i.e. without gene selection) cannot predict outcome at all ( $p > 0.1$ ).

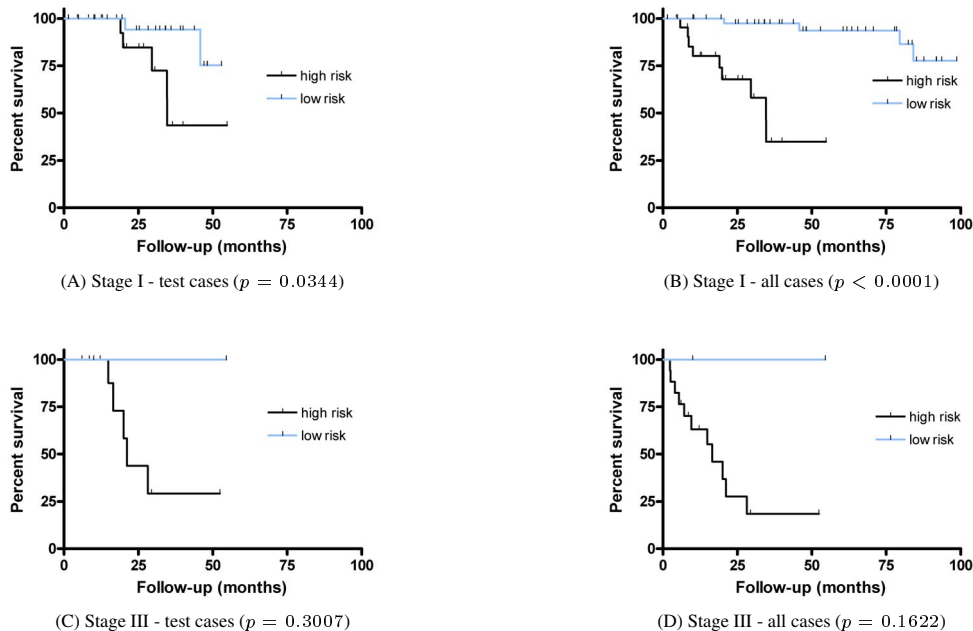
In the step of prediction, a simple linear kernel SVM is trained on the selected samples and genes to build a scoring model. The model then assigns each validation sample a risk score to predict patient outcome. Based on the training results, we can derive explicit thresholds (e.g., 0.5, 0.3, 0.7) of 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.

For both studies on DLBCL and lung adenocarcinoma,

the results that we obtained illustrate that samples assigned to the high and low risk groups based on our score have significant difference on survival. Besides, in DLBCL study, our high and low risk groups also demonstrated significantly different outcomes in the analysis of patients with low or intermediate risk according to their International Prognostic Index (IPI) scores constructed on some clinical features. E.g., for patients having high IPI, we assign most of them into our high risk category and some of them into our low risk category, and our assignment is better correlated to survival outcome of these patients. Some of the genes identified to have strong association with survival by our filtering method also fall within four biologic groups defined on the basis of gene expression signatures. In the lung adenocarcinoma study, most of the samples are from stage



**Figure 4. Kaplan-Meier plots illustrate the estimation of overall survival among high risk and low risk lung adenocarcinoma patients in the testing set containing 55 samples (Panel (A)) and all 86 samples (Panel (B)).**



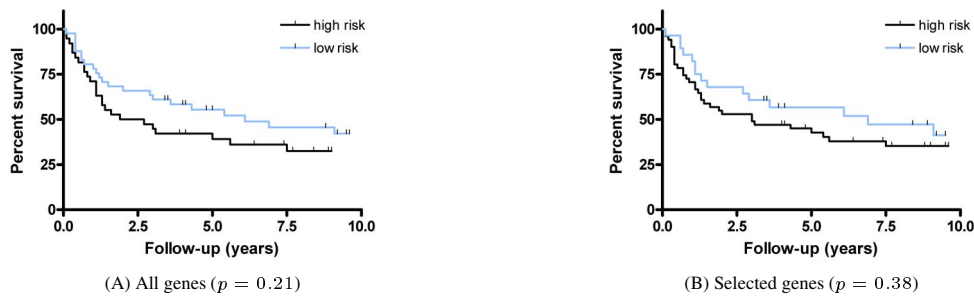
**Figure 5. Kaplan-Meier plots illustrate the estimation of overall survival among high risk and low risk lung adenocarcinoma patients conditional on tumor stage.**

I tumors. Among these samples, although our high and low risk groups differ significantly from each other, we put quite a few of them into high risk group. This finding “*indicates the important relationship between gene expression profiles and patient survival, independent of disease stage*”, which is one of the conclusions drawn in [3].

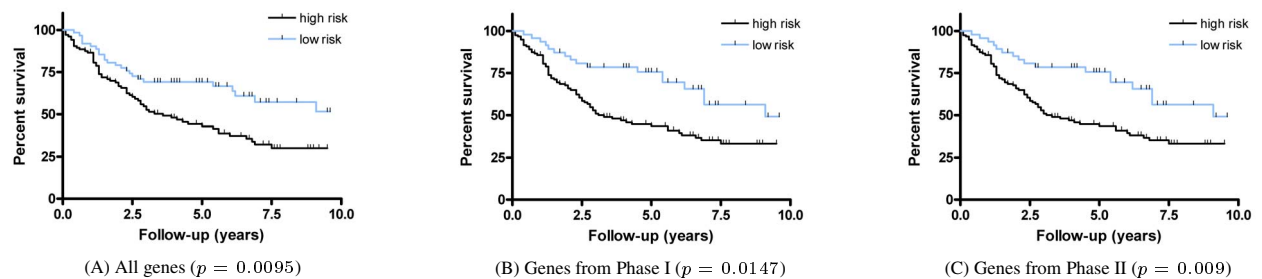
In summary, we have applied statistical and machine learning technologies to predict patient outcome using gene expression profiles. Different from other works, we first pick out extreme cases to form the training set, consisting

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. Naturally, if there are indeed genes associated with outcome, then the different expression values of these genes should be monitored by analysing these two types of samples.

We have some ongoing works: (1) data sets from other tumors are under analysis; and (2) for some particular diseases, we are further extracting biological meanings of the genes identified to be most associated with patient survival.



**Figure 6.** Kaplan-Meier plots illustrate no clear difference on the overall survival among high risk and low risk DLBCL patients formed by the 80 validation samples based on their risk scores that assigned by our regression model built on all 160 training samples. (A) Using all genes. (B) Using genes selected by our method.



**Figure 7.** Kaplan-Meier plots illustrate the estimation of overall survival among 167 high risk and low risk testing samples in DLBCL study. (A) Using all 4937 genes. (B) Using 132 genes output from the Phase I filtering. (C) Using 84 genes output from the Phase II filtering.

| Gene selection | DLBCL      | Lung adenocarcinoma |
|----------------|------------|---------------------|
| Original       | 4937(*)    | 7129                |
| Phase I        | 132 (2.7%) | 884 (12.4%)         |
| Phase II       | 84 (1.7%)  | 591 (8.3%)          |

**Table 2.** Number of genes left after feature filtering for each phase. The percentage in the brackets indicates the proportion of the remaining genes on original feature space. (\*)The number is after removing genes who were absent in more than 10% of the experiments.

## Appendix

### A: Entropy measure

Using entropy measure for feature selection is inspired by the algorithm of Fayyad and Irani [6] for discretizing

numeric features in classification problems. This discretization algorithm finds some cut point(s) for a numeric feature's value range to make the resulting value intervals as pure as possible. For those features whose values are relatively randomly distributed between different class of samples, the algorithm will not find any proper cut point, and we should thus discard them.

Formally, given a collection  $S$ , containing samples in  $k$  classes, the *entropy* of  $S$  relative to this  $k$  classes classification is defined as

$$Ent(S) \equiv \sum_{i=1}^k -p_i * \log_2 p_i \quad (6)$$

where  $p_i$  is the proportion of  $S$  belonging to class  $i$ . Then Fayyad's discretization algorithm works as follows. Let cut point  $T$  of feature  $X$  partition the sample set  $S$  into the subsets  $S_1$  and  $S_2$ . The *class information entropy* of the parti-

tion, denoted  $Ent(X, T; S)$ , is given by [6]:

$$Ent(X, T; S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2) \quad (7)$$

where  $Ent(S_j)$ , ( $j = 1, 2$ ) is the class entropy of a subset  $S$ . Suppose there are  $k$  classes  $C_1, \dots, C_k$ . Let  $P(C_i, S_j)$  be the proportion of samples in  $S_j$  that have class  $C_i$ . According to the definition in Equation (6),

$$Ent(S_j) = - \sum_{i=1}^k P(C_i, S_j) * \log_2(P(C_i, S_j)) \quad (8)$$

A binary discretization for  $X$  is determined by selecting the cut point  $T_X$  for which  $E(X, T; S)$  is minimal amongst all the candidate cut point [6]. This can be achieved by recursively partitioning the ranges  $S_1$  and  $S_2$  until some stopping criteria is reached. In this study, we adopt the commonly used stopping criteria called *minimal description length (MDL) principle* [6]. Generally, under the entropy measure, feature  $X$  is more useful than feature  $Y$  if  $Ent(X, T_X; S) < Ent(Y, T_Y; S)$ .  $T_X, T_Y$  is the best cut point found for  $X$  and  $Y$ , respectively.

## B: Wilcoxon rank sum test

Wilcoxon rank sum test, or the equivalent Mann-Whitney test, is a kind of non-parametric test since it is based on rank of samples rather than distribution parameters such as mean and standard deviation. It does not require the two populations to conform to a normal distribution, but the same shape [12].

If a data set with two classes  $C_1$  and  $C_2$  is presented by a matrix  $D = \{x_{ij}\}$ , where  $1 \leq i \leq n$ ,  $1 \leq j \leq m$ ,  $n$  is the number of samples, and  $m$  is the number of features, the Wilcoxon rank sum test statistical measure of a feature  $X_j$ ,  $w(X_j)$ , can be obtained using following procedure:

- Sort the values  $x_{1j}, x_{2j}, \dots, x_{nj}$  of  $X_j$  across all the samples in an ascending order.
- Assign rank (from 1)  $r(x_{ij})$  to each value  $x_{ij}$  above and use average of the ranks for ties. Then,  $1 \leq r(x_{ij}) \leq n$ .
- Use the sum of the ranks for the class, which has smaller number of samples, as test statistic,  $w(X_j)$ . E.g., class  $C_1$  has fewer samples than class  $C_2$ , then

$$w(X_j) = \sum_{i \in C_1} r(x_{ij}) \quad (9)$$

If the number of samples is same in each class, the choice of which class to use for the test statistic is arbitrary.

To use the Wilcoxon rank sum test to decide if a feature  $X$  is relevant, we set up the null hypothesis that: the values of  $X$  are not much different in  $C_1$  and  $C_2$ . Then  $w(X)$  is used to accept or reject the hypothesis. To decide whether to accept or reject the null hypothesis, we compare  $w(X)$  with the upper and lower critical values derived from a significant level  $\alpha$ . For smaller numbers of samples in class  $C_1$  ( $n^{C_1}$ ) and  $C_2$  ( $n^{C_2}$ ), e.g.  $< 10$ , the critical values have been tabulated and can be found in [12]. If either  $n^{C_1}$  or  $n^{C_2}$  is larger than what is supplied in the table, the following normal approximation can be used [12]. The expected value of  $w$  is:

$$\mu = \frac{n^{C_1} * (n^{C_1} + n^{C_2} + 1)}{2} \quad (10)$$

assuming class  $C_1$  has fewer samples than class  $C_2$  does. The standard deviation of  $w$  is:

$$\sigma = \sqrt{\frac{n^{C_1} * n^{C_2} * (n^{C_1} + n^{C_2} + 1)}{12}} \quad (11)$$

The formula for calculating the upper and lower critical values is:

$$\mu \pm z_\alpha \sigma \quad (12)$$

where  $z_\alpha$  is the  $z$  score for significant level  $\alpha$ . If a feature  $X$ 's test  $w(X)$  falls in the range given by the upper and lower critical values, then we accept the null hypothesis; otherwise, reject the hypothesis, and this indicates that the values of feature  $x$  are significantly different between samples in class  $C_1$  and  $C_2$ .

## References

- [1] Altman, D.B. *Practical statistics for medical research*, Chapman and Hall, 1991.
- [2] Ando, T., & Katayama, M., 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, 2002.
- [3] Beer, D.G., Kardia, S.L., Huang, C.C., Giordano, T.J., Levin, A.M., Misek, D.E., Lin, L., Chen, G., Gharib, T.G., Thomas, D.G., Lizyness, M.L., Kuick, R., Hayasaka, S., Taylor, J.M., Iannettoni, M.D., Orringer, M.B. & Hanash, S., Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**(8):816–823, 2002.
- [4] Borczuk, A.C., Gorenstein, L., Walter, K.L., Assaad, A.A., Wang, L., & Powell, C.A., Non-small-cell lung cancer molecular signatures recapitulate lung developmental pathways. *Am. J. Pathol.*, **163**(5):1949–1960, 2003.
- [5] Cox, D.R., Regression models and life-tables (with discussion). *J. R. Stat. Soc.*, **B34**:187–220, 1972.
- [6] Fayyad, U. & Irani, K., Multi-interval discretization of continuous-valued attributes for classification learning. In *13th Intl. Joint Conf. Artificial Intelligence*, pp. 1022–1029, 1993.

- [7] LeBlanc, M., Kooperberg, C., Grogan, T.M., & Miller, T.P., Directed indices for exploring gene expression data. *Bioinformatics*, **19**(6):686–693, 2003.
- [8] Li, J., Liu, H., & Wong, L., Mean-entropy discretized features are effective for classifying high-dimensional biomedical data. *3rd ACM SIGKDD Workshop on Data Mining*, pp. 17–24, 2003.
- [9] Lunn, M., & McNeil, D.R., Applying Cox Regression to Competing Risks. *Biometrics*, **51**:524–532, 1995
- [10] Park, P.J., Tian, L., & Kohane, S., Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18**(Suppl 1):S120–S127, 2002.
- [11] Rosenwald, A. *et al*, The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma, *NEJM*, **346**(25):1937–1947, 2002.
- [12] Sandy, R., *Statistics for Business and Economics*. McGrawHill, 1989.
- [13] Shipp, M.A. *et al*, A predictive model for aggressive non-Hodgkin’s Lymphoma. *NEJM*, **329**:987–994, 1993.
- [14] Shipp, M.A. *et al*, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, **8**(1):68–74, 2002.
- [15] Vapnik, V.N., *The Natural of Statistical Learning Theory*, Springer-Verlag New York Inc., 1995.
- [16] van de Vijver, M.J. *et al*, A gene-expression signature as a predictor of survival in breast cancer. *NEJM*, **347**(25):1999–2009, 2002.
- [17] Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics*, **1**, 80–83, 1945.
- [18] Witten, H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*. Morgan Kaufmann, 2000.
- [19] Yang, Y., Mou, L.j., Liu, N., & Tsao, M.S., Autotaxin in expression in non-small-cell lung cancer. *Am. J. Respir. Cell Mol. Biol.*, **21**(2):216–222, 1999.
- [20] Yeoh, E.-J. *et al*, Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**:133–143, 2002.