Graduate Theses, Dissertations, and Problem Reports

2008

Constructing gene expression based prognostic models to predict recurrence and lymph node metastasis in colon cancer

Ramakanth Reddy Mettu
West Virginia University

Constructing gene expression based prognostic models to predict recurrence and lymph node metastasis in colon cancer

Ramakanth Reddy Mettu

Thesis Submitted to the

College of Engineering and Mineral Resources

at West Virginia University

in partial fulfillment of the requirements

for the degree of

Master of Science

in

Electrical Engineering

Nancy Lan Guo, Ph.D., Chair

Bojan Cukic, Ph.D.

Tim Menzies, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia

2008

Keywords: Random forests, Feature selection, Machine learning, Classification, Colon cancer, Recurrence, Kaplan-Meier

ABSTRACT

Constructing gene expression based prognostic models to predict recurrence and lymph node metastasis in colon cancer

Ramakanth Reddy Mettu

The main goal of this study is to identify molecular signatures to predict lymph node metastases and recurrence in colon cancer patients. Recent advances in microarray technology facilitated building of accurate molecular classifiers, and in depth understanding of disease mechanisms.

Lymph node metastasis cannot be accurately estimated by morphological assessment. Molecular markers have the potential to improve prognostic accuracy. The first part of our study presents a novel technique to identify molecular markers for predicting stage of the disease based on microarray gene expression data. In the first step, random forests were used for variable selection and a 14-gene signature was identified. In the second step, the genes without differential expression in lymph node negative versus positive tumors were removed from the 14-gene signature, leading to the identification of a 9-gene signature. The lymph node status prediction accuracy of the 9-gene signature on an independent colon cancer dataset (*n=17*) was 82.3%. Area under curve (AUC) obtained from the time-dependent ROC curves using the 9-gene signature was 0.85 and 0.86 for relapse-free survival and overall survival, respectively. The 9-gene signature significantly stratified patients into low-risk and high-risk groups (log-rank tests, *p<0.05, n=73*), with distinct relapse-free survival and overall survival. Based on the results, it could be concluded that the 9-gene signature could be used to identify lymph node metastases in patients. We further studied the 9-gene signature using correlation analysis on CGH and RNA expression datasets. It was found that the gene *ITGB1* in the 9-gene signature exhibited strong relationship of DNA copy number and gene expression. Furthermore, genome-wide correlation analysis was done on CGH and RNA data, and three or more consecutive genes with significant correlation of DNA copy number and RNA expression were identified. These results might be helpful in identifying the regulators of gene expression.

The second part of the study was focused on identifying molecular signatures for patients at high-risk for recurrence who would benefit from adjuvant chemotherapy. The training set (*n=36*) consisted of patients who remained disease-free for 5 years and patients who experienced recurrence within 5 years. The remaining patients formed the testing set (*n=37*). A combinatorial scheme was developed to identify gene signatures predicting colon cancer recurrence. In the first step, preprocessing was done to discard undifferentiated genes and missing values were replaced with *k*=30 and *k*=20 using the *k*-nearest neighbors algorithm. Variable selection using the random forests algorithm was applied to obtain gene subsets. In the second step, InfoGain feature selection technique was used to drop lower ranked genes from the gene subsets based on their association with disease outcome. A 3-gene and a 5-gene signature were identified by this technique based on different missing value replacement methods. Both of the recurrence gene signatures stratified patients into low-risk and high-risk groups (log-rank tests, *p<0.05, n=73*), with distinct relapse-free survival and overall survival. A recurrence prediction model was built using LWL classifier based on the 3-gene signature with an accuracy of 91.7% on the training set *(n=36)*. Another recurrence prediction model was built using the random tree classifier based on the 5-gene signature with an accuracy of 83.3% on the training set *(n=36)*. The prospective predictions obtained on the testing set using these models will be verified when the follow-up information becomes available in the future. The recurrence prediction accuracies of these gene signatures on independent colon cancer datasets were in the range 72.4% to 88.9%. These prognostic models might be helpful to clinicians in selecting more appropriate treatments for patients who are at high-risk of developing recurrence. When compared over multiple datasets, the 3-gene signature had improved prediction accuracy over the 5-gene signature. The identified lymph node and recurrence gene signatures were validated on rectal cancer data. Time-dependent ROC and Kaplan-Meier analysis were done producing significant results. These results support the fact that the developed prognostic models could be used to identify patients at high-risk of developing recurrence and get an estimate of the survival times in rectal cancer patients.

Acknowledgements

I take this opportunity to express my gratitude and appreciation to my advisor, Dr. Nancy Lan Guo, for her continual guidance and support throughout my work in this research. Dr.Guo has been very patient and kind towards me. I had the opportunity to learn a lot of new things while working on this project. This research project was supported by the grant NIH/NCRR P20 RR16440-03.

I would like to thank Dr. Menzies and Dr. Cukic, my committee members, for their suggestions and guidance. I would like to thank Dr.Thomas Ried for providing us with the data.

I want to express my sincere thanks to Dr. Ross for teaching the Pattern Recognition course which helped me a lot in this research.

I would like to thank Dr.Yan Ma and Zhenyu Ding for teaching me the required skills and helping me to get started with the research. I would also like to thank my lab members Shruti Rathnagiriswaran, Kursad Tosun, Ying Wooi Wan and Swetha Nutakki for their help and feedback on my research.

I appreciate the help of Suhasini Kalluru in revising this thesis. I am grateful to my family and friends for all the help and support they have extended without which I couldn’t have gone this far.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Colon cancer is the third most common cause of cancer in Europe and the United States, with ~300,000 new cases and 200,000 deaths each year. It is the second most common site (after lung) to cause cancer death[1]. The primary treatment for colon cancer is the surgical removal of a part of colon or the entire colon. Chemotherapy after surgery can prolong the survival in patients if the cancer has spread to nearby lymph nodes. Prognosis is the estimation of disease outcome i.e., the chance that a patient will recover or have a recurrence (return of the cancer)[2]. The most important factors that affect the colon cancer prognosis are the histology, location, and stage of the disease (the extent to which the cancer has spread). Doctors cannot be absolutely certain about the outcome for a particular patient based on the traditional morphological assessment.

In the recent years, advances in genetic technologies such as cDNA microarrays allowed for measuring the expression of tens of thousands of genes simultaneously. The research carried out in this area over the past few years has demonstrated that the gene expression data could be used to solve a variety of problems like tumor classification and prediction of treatment response. Machine learning and statistical techniques have been successfully applied on the gene expression datasets to identify biomarkers, predict recurrence or disease outcome, distinguish between tumor and normal tissue samples, build prognostic predictors and predict treatment response (1). Currently, two gene expression based tests are being used in clinical trials for breast cancer prognosis. The MammaPrint[3] test classifies tumors into

[1] http://en.wikipedia.org/wiki/Colorectal_cancer

[2] http://colon-cancer.emedtv.com/colon-cancer/colon-cancer-prognosis.html

[3] http://usa.agendia.com/en/mammaprint.html

low or high-risk of recurrence. The Oncotype DX[4] test determines the likelihood of recurrence. There are no gene tests available for colon cancer prognosis at present.

Staging is an important prognostic factor in determining treatment options. Earlier stages of colon cancer (stage I and stage II) have good chances of prognosis compared to later stages (stage III and stage IV)[5]. When a patient is diagnosed with cancer, various clinical parameters are used to assess the risk of metastasis and death in the patient. However, despite numerous advances in this area, the ability to accurately estimate the risk of morbidity is limited. Tumors that appear indistinguishable under the microscope can have different outcome and different treatment response. This could be due to the differences in the genetic profiles of the tumors. With the advent of cDNA microarray technology, it is possible to measure the expression levels of thousands of genes simultaneously and the differences between tumors at the molecular level can be detected. Thus molecular markers identified based on the cDNA microarray expression data have the potential to improve prognostic accuracy significantly. The disease prognosis can be assessed preoperatively through a tissue obtained from a colonoscopic biopsy specimen or post operatively from a resected tumor.

The first part of our study aims at building prognostic models based on the microarray gene expression data to predict lymph node metastasis (stage). The colon cancer microarray data used in this study contained 73 tumor samples of which 33 samples were stage II tumors and 40 samples were stage III tumors (2). The data was preprocessed by applying t-tests[6] on genes that had missing values in more than 5 samples. Genes passing the t-tests along with all genes having less than 5 missing values, a total of 10,220 genes, were included for further analysis. This data was randomly split in 2:1 ratio as training and testing sets. The training set contained 10,220 genes and 50 samples. A 9-gene lymph node status signature was identified by a novel technique from the training set. In this technique, firstly, variable selection using random forests (3) was done and a 14-gene signature was identified. In the next step, z-

[4] http://www.genomichealth.com/oncotype/default.aspx

[5] http://medicineworld.org/cancer/colon/colon-cancer-staging.html

[6] http://www.socialresearchmethods.net/kb/stat_t.php

tests[7] were applied on the 14-gene signature to discard the genes without differential expression in lymph node negative versus positive tumor samples. This resulted in a 9-gene signature. The performance of the 9-gene signature was evaluated by cross validation on independent colon cancer data sets. A number of machine learning algorithms were applied on validation datasets, but none of the algorithms gave consistent results on all the validation datasets. So, classifiers with highest prediction accuracy on each dataset were chosen. *J48*, Naïve Bayes, Decision stump and Threshold selector were the classifiers used for validation of independent datasets. The 9-gene signature was used to predict lymph node status on Koinuma et al data ($n=17$), recurrence on Barrier et al data (PMID 16091735) ($n=12$), Barrier et al data (PMID 16966692) ($n=50$), Barrier et al data (PMID 17043639) ($n=24$), and drug response on NCI-60[8] data ($n=34$). Further time-dependent ROC analysis was done to get an estimate of the discriminatory power of the identified biomarker. Kaplan-Meier analysis generated significant patient stratification into subgroups ($p<0.05$, $n=73$, log-rank tests) with distinct relapse-free survival and overall survival, respectively. Correlation analysis was done on CGH and RNA data to identify cDNA copy numbers correlated with gene expression data. The locations of these genes might be important in identifying the regulators of the gene expression (4).

Recurrence is the reappearance of a tumor or the return of symptoms after treating for cancer. Adjuvant chemotherapy is the main treatment given to Duke’s stage C patients (node-positive disease). In Duke’s stage B patients (node negative disease) no adjuvant chemotherapy is used after surgery, although 25% to 40% of patients usually develop recurrence (5). It is not clear whether adjuvant chemotherapy should be given to Duke’s stage B patients as not all the patients would benefit from it. Partitioning patients into low-risk and high-risk groups would allow in “more aggressive” and accurate treatment strategies for the patients at high-risk of recurrence, and spare the patients in the low-risk group from the “aggressive treatment” through which they are unlikely to be benefited. The TNM (tumor-node-metastasis) staging system is the main tool for identifying prognostic differences (6), but this system is

[7] http://en.wikipedia.org/wiki/Z-test

[8] http://discover.nci.nih.gov/cellminer/loadDownload.do

not sufficient for predicting recurrence in Duke's stage B patients (7). Thus, there are limitations for predicting recurrence by using traditional methods. So there is a need to identify patients at high-risk of recurrence who would develop relapse in the Duke's B group.

The second part of our study specifically aims at identifying patients at high-risk of recurrence by building prognostic models for stage II (Duke's stage B) and stage III (Duke's stage C) colon cancer patients. This is achieved by a novel combinatorial feature selection scheme. The missing values in the gene expression data were replaced by *k*-nearest neighbors algorithm with *k*=30 and *k*=20, separately. In the first step, variable selection using random forests is done on the training set which comprised of 36 patients. This step obtained two recurrence gene signatures based on different missing value replacement methods. In the second step, InfoGain feature selection technique (12) was applied to further reduce the dimensionality, and this led to the identification of the 3-gene signature and the 5-gene signature on datasets generated with different missing value replacements. The performances of both gene signatures were evaluated by cross validation on independent colon cancer data sets. A number of machine learning algorithms have been tested for the validation of these signatures, but no particular scheme gave consistent results on all the datasets. So, classifiers with highest prediction accuracy on each dataset were chosen. LWL and Random Tree were the classifiers chosen to build prediction models using 3-gene and 5-gene signatures, respectively. *KStar*, AD Tree, *IB1* and Threshold selector were the classifiers used for validation of independent datasets. The 3-gene and 5-gene recurrence signatures were used to predict lymph node status on Koinuma et al data (*n=17*), recurrence on Barrier et al data (PMID 16091735) (*n=12*), Barrier et al data (PMID 16966692) (*n=50*), Barrier et al data (PMID 17043639) (*n=24*), and drug response on NCI-60[9] data (*n=34*) independently. Further, time-dependent ROC analysis was done to get an estimate of the discriminatory powers of the identified gene signatures. Prediction models were built with the 3-gene signature and the 5-gene signature using classifiers in Weka[10] to predict recurrence in patients from the testing set. Kaplan-Meier analysis using the 3-gene signature generated significant

[9] http://discover.nci.nih.gov/cellminer/loadDownload.do

[10] http://www.cs.waikato.ac.nz/ml/weka/

patient stratification into low-risk and high-risk groups (*$p < 0.05$, $n=73$*, log-rank tests,) with distinct relapse-free survival and overall survival, respectively. Kaplan-Meier analysis using the 5-gene signature generated significant patient stratification into low-risk and high-risk groups (*$p < 0.05$, $n=73$*, log-rank tests) with distinct relapse-free survival and overall survival, respectively. When the 3-gene and 5-gene signatures were compared over multiple datasets the 3-gene signature had improved prediction accuracy. But the difference in the prediction accuracies was not statistically significant. From these results it can be concluded that it is possible to build prognostic models based on the microarray gene expression data to identify patients at high-risk of recurrence. The identified gene signatures were validated on rectal cancer data and they generated significant patient stratification into low-risk and high-risk groups.

This thesis is organized as follows. Chapter 2 discusses the background of our study. Chapter 3 describes the experimental details of the identification and validation of the 9-gene signature and the validation results. Chapter 4 describes the experimental details of identification and validation of the 3-gene and 5-gene recurrence signatures. Chapter 5 discusses the validation results of all the gene signatures on rectal cancer data, and Chapter 6 concludes this study.

Chapter 2

Background

2.1 Introduction

Gene signatures can be used to aid clinical decision-making in personalized therapy. They can also be used to stratify patients who would experience recurrence and who would not. The goal of our study is to identify a small subset of genes that could potentially be used to predict the likelihood of lymph node metastases (stage) and recurrence in patients with colon cancer. Prognostic models can be built based on these gene signatures to identify patients at high-risk of recurrence. These gene signatures have the potential for improving diagnostic classification, treatment selection, and prognostic assessment.

The advent of high-throughput technologies such as DNA microarrays is currently revolutionizing biology and medicine. Machine learning techniques are playing a pivotal role in analyzing the generated microarray data. Machine learning algorithms are very useful in cancer research and several machine learning algorithms have already been successfully applied on microarray gene expression data to classify tumors, predict disease outcome and treatment response (8). Unsupervised machine learning approaches such as, self-organizing maps (SOM) were used to organize genes into biologically relevant clusters in leukemia (11), and hierarchical clustering was used to classify colon cancer tissues into cancerous and non-cancerous based on the gene expression (9). Supervised machine learning techniques such as Support vector machines (SVMs) were used for multi-class cancer diagnosis (10). Nearest shrunken centroids were used for diagnosing cancer (38). Decision trees and feed-forward neural networks were used for lung cancer classification (39).

The remainder of this chapter is organized as follows. Section 2.2 describes the feature selection techniques utilized in this study. Section 2.3 describes the classification algorithms used in this study. Section 2.4 explains the survival analysis techniques. Section 2.5 describes the correlation coefficient analysis. Section 2.6 presents the related work performed in previous studies. Section 2.7 discusses the open problems in this area, and Section 2.8 summarizes this chapter.

2.2 Feature selection techniques

Two of the most important problems in microarray data analysis relate to the dimensionality of the data and noise. In many bioinformatics problems, the number of features is significantly larger than the number of samples (high feature to sample ratio). Moreover, not all the features are necessary for classification purposes. Inclusion of all the features would contribute noise and introduce an error.

Feature selection is the process of systematically reducing the dimensionality of a dataset to an optimal subset of attributes for classification purposes. The main idea of feature selection is to choose a subset of input variables. Feature selection can significantly improve the comprehensibility of the resulting classifier models by eliminating features with little or no predictive information. Several commonly used feature selection techniques like Random forests, Information gain attribute evaluator, CfsSubset evaluator, GainRatio evaluator, and ReliefF attribute evaluator are described as follows.

2.2.1 Variable selection using Random forests

Random forests are an ensemble method that combines several individual classification trees. In order to grow these ensembles, random vectors are generated that govern the growth of each tree in the ensemble. The basic step of random forests is to form diverse tree classifiers from a single training set. Each tree is built upon a “bootstrap sample” taken from the training set. A random subset from the whole set of variables are used for splitting the tree nodes. The classification decision of a new case is obtained by majority voting over all trees unless the cut-off value is user defined. In random forests, about one-third of the cases in the bootstrap sample are not used in growing the tree. These cases are called “out-of-bag” (OOB) cases and are used in evaluating the performance of the algorithm.

Random forest returns several measures of variable importance. The most reliable measure is the “mean decrease in accuracy”. Mean decrease in accuracy considers the importance of an m^{th} variable as the difference between the “out-of-bag” error rate for the randomly permuted m^{th} variable (the error rate obtained by randomly rearranging the values of the m^{th} variable for the out-of-bag set, for each tree, and getting new classifications for the forest, by putting this permuted set down the tree) and the original “out-of-bag” error rate (41). Based on the “mean decrease in accuracy” measure, backward elimination was used to identify the gene subset with the smallest “out-of-bag” error rate. The OOB error rate was used to choose the final set of genes, not to obtain estimates of the error rate. This procedure was implemented using the varSelRF[11] package in R[12] software.

2.2.2 Information gain attribute evaluator

Information gain (InfoGain) attribute evaluator is a supervised attribute filter for selecting attributes. This method evaluates the attributes by measuring information gain with respect to class. Numeric attributes are first discretized using the MDL-based discretization method[13]. This method treats missing value as a separate value or distributes the counts among other values in proportion to their frequency. It is used in conjunction with the Ranker which ranks attributes by their individual evaluations. It is only capable of generating attribute rankings (12). The user can specify the number of attributes to retain and the threshold can be adjusted to discard the attributes.

The information gain of a given attribute X with respect to the class attribute Y is given by:

$$I(Y;X) = H(X) + H(Y) - H(X,Y) \qquad \textit{(Equation 1)}$$

where $H(X)$ is the entropy of X, $H(Y)$ is the entropy of Y, and $H(X,Y)$ is the joint entropy of X and Y.

2.2.3 CfsSubset evaluator

Subset evaluators take a subset of attributes and return a numeric measure that guides the search. CfsSubset evauator assesses the predictive ability of each attribute individually and the degree of

[11] http://cran.r-project.org/web/packages/varSelRF/index.html

[12] http://www.r-project.org/

[13] http://de.scientificcommons.org/20784480

redundancy among them, preferring sets of attributes that are highly correlated with the class but having low inter-correlation. Conditional entropy is used to provide a measure of the correlation between features and class and between features. If *H(X)* is the entropy of a feature *X* and *H(X|Y)* the entropy of a feature *X* given the occurrence of feature *Y,* the correlation between two features *X* and *Y* can then be calculated using the symmetrical uncertainty as follows:

$$C(X/Y) = \frac{H(X)-H(X/Y)}{H(Y)} \qquad (\textit{Equation 2})$$

The class of an instance is considered to be a feature. The goodness of a subset is then determined as:

$$G_{subset} = \frac{kr_{ci}}{\sqrt{(k}+k(k-1)r_{ii})} \qquad \textit{(Equation 3)}$$

where *k* is the number of features in a subset, r_{ci} is the mean feature correlation with the class and r_{ii} is the mean feature correlation.

2.2.4 GainRatio attribute evaluator

GainRatio attribute evaluator evaluates attributes by measuring their gain ratio with respect to the class. If *X* represents the attribute and *Y* represents the class the GainRatio is given by the following equation:

$$GainR(Y,X) = \frac{H(Y)-H(Y/X)}{H(X)} \qquad \textit{(Equation 4)}$$

where *H(Y)* is the entropy of *Y*, *H(X)* is the entropy of *X*, and *H(Y/X)* is the entropy of *Y* given *X*.

Missing value counts can be distributed across other values in proportion to their frequency or they can be treated as separate values.

2.2.5 ReliefF attribute evaluator

ReliefF attribute evaluator evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class. It can operate on both discrete and continuous class data. *ReliefF* generalizes the behavior of Relief to classification. It finds one nearest neighbor of I_1 from every class. On these neighbors Relief evaluates the relevance of every feature $f \in F$ accumulating it into *W[f]*. The nearest neighbor from the same class is a hit *H*, and from a different class is a miss *M(C)* of class *C*. At the end *W[f]* is divided by *m* to get the average evaluation in [–1, 1].

$$W[f] = W[f] - diff(f, I_1, H) + \sum_{C \neq class(R_i)} [\, P(C) * diff(f, I_1, M(C))] \qquad (Equation\ 5)$$

The function *diff* $(f; I_1; I_2)$ calculates the difference between the values of the attribute *A* for two instances I_1 and I_2. For nominal attributes it is defined as:

$$diff(f, I_1, I_2) = \begin{cases} 0; & value(f, I_1) = value(f, I_2) \\ 1; & otherwise \end{cases} \qquad (Equation\ 6)$$

For numerical attributes it is defined as:

$$diff(f, I_1, I_2) = \frac{|value(f, I_1) - value(f, I_2)|}{\max(f) - \min(f)} \qquad (Equation\ 7)$$

2.3 Classification algorithms

Machine learning is a subfield of Artificial Intelligence dealing with the development of algorithms that learn from past experience. Machine learning techniques are extensively applied to microarray data, particularly for diagnostic purposes. Especially in cancer diagnostics, microarray classification tools are used for cancer subtype discrimination and outcome prediction. The following section describes the machine learning algorithms that we have used in our research for predicting disease subtype and outcome, and building prognostic models.

2.3.1 Bagging

Bagging stands for bootstrap aggregating. Given a training set, the original training data is altered by deleting some instances and replicating others. Instances are randomly sampled with replacement from the original dataset to create a new one of the same size. Instead of obtaining independent datasets from the domain, bagging just resamples the original training data. Then, a learning scheme like a decision tree is applied to each of these derived datasets and the classifiers generated from them vote for the class to be predicted. All models receive equal weights and bagging produces a combined model that often performs significantly better than the single model built on the original training data (12).

2.3.2 Naive Bayes

The classifier is named so, because it is based on Baye's rule and assumes that the attributes are independent "naively". It is particularly suitable when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. If the data is redundant, Naive Bayes classifier works well with some attribute selection procedures that eliminate redundant data. The Bayes rule is described as follows.

If H is the hypothesis and E is the evidence that bears on that hypothesis, then

$$P\,(H|E) = P(E|H)\,P(H)\,/\,P(E) \qquad \textit{(Equation 8)}$$

$$\text{or} \quad Posterior = \frac{Likelihood * Prior}{Evidence} \qquad \textit{(Equation 9)}$$

2.3.3 Threshold selector

Threshold selector is a Meta classifier that selects a threshold on the probability distribution output by a classifier. The threshold is set so that a given performance measure is optimized. The performance measure is the F-measure[14] (Equation 3). Performance can be measured either on the training data, on a hold-out set, or using cross-validation (12).

$$F - measure = \frac{2 * True\ positive}{2 * True\ positive + False\ positive + False\ negative} \qquad \textit{(Equation 10)}$$

[14] http://en.wikipedia.org/wiki/Information_retrieval

2.3.4 Locally Weighted Learning (LWL)

LWL belongs to the class of instance-based learners. It assigns weights using an instance-based method and builds a classifier from the weighted instances. Attribute normalization is turned on by default. The base classifier can be selected by the user. Naive Bayes is a good choice for classification problems. Other parameters that can be adjusted are *k*-nearest neighbor (KNN). This method determines the number of neighbors used to determine the width of the weighting function, and the kernel shape to use for weighting, which can be linear, inverse, constant or Gaussian (12).

2.3.5 Multilayer Perceptron

Multilayer Perceptron is a neural network classifier. It belongs to the class of supervised neural networks. It is one of the most important and widely used network models**.** The multi-layer perceptron neural network model consists of a network of processing elements or nodes arranged in layers, usually interconnected in a feed-forward way. Each neuron in one layer has directed connections to the neurons of the subsequent layer. This classifier uses back propagation technique for learning. In MLPs, learning is supervised with separate training and recall phases.

1. The network produces an output pattern for each input pattern.
2. The actual output is compared with the known output from the training set and the error is calculated.
3. The weights are adjusted to reduce the error.
4. The steps 1-3 are repeated many times for every instance in the training set until the error is minimized.

Once the network has been trained, the weights are then fixed. The testing set is fed into the network and the network output is compared with the desired output.

2.3.6 J48

The Weka package implements its own version of C4.5 known as *J48*. This algorithm induces decision trees for classification by using the greedy technique. A decision-tree model is built by analyzing the training data and that model is used to classify testing data. If the test data is not available, *J48* performs a cross-validation using the training data.

2.3.7 IB1

The *IB1* classifier is a 1-nearest neighbor instance-based classifier. It is the simplest instance-based learning algorithm. It uses a simple distance measure to find the training instance closest to the given test instance and assigns the same class as that of the training instance. If multiple closest instances are found, the first one found is used. Generally the distance measure used is the Euclidean distance. An advantage of instance-based learning over many other machine learning methods is that new examples can be added to the training set at any time. Though instance-based learning is simple and works very well, it is often slow (12).

2.3.8 KStar

KStar is an instance-based classifier, meaning that the class of a test instance is based upon the class of those training instance(s) that resemble it most. The resemblance is calculated by using the distance function. *KStar* uses an entropy-based distance function. This way it differs from other instance-based classifiers. It belongs to the class of *k*-nearest neighbor classifiers because it classifies each instance by looking at the nearest *k* data points and determining the class by the one which is the most common in the nearest *k* data points (13). *KStar* has an option to specify the blend factor which specifies how the distance function used to compute the *k*-nearest neighbors acts. If the blend factor is set to 0%, the distance function performs like a standard nearest neighbor classifier by selecting just one instance to classify the test instance. If the blend factor is set to 100%, the distance function takes many instances and then classifies by the most common class.

2.3.9 Alternating Decision Tree (AD Tree)

Alternating decision tree is a generalized representation of both voted stumps and decision trees. It uses boosting as a method for learning data. AD Tree supports only two-class problems. The number of boosting iterations can be manually tuned to suit the dataset and the desired complexity/accuracy tradeoff. More boosting iterations result in larger and potentially more accurate trees, but make the learning process slower (12). Each of the iterations adds three nodes to the tree (one split node and two prediction nodes) unless merging occurs. The default search method is an exhaustive search. Heuristic search methods can be used to speed up learning but they are not guaranteed to find an optimal solution. The instance data can be saved for visualization.

2.3.10 AdaboostM1

AdaboostM1 is a variant of Adaboost technique for multi-class problems. Adaboost stands for adaptive boosting. Boosting is one type of meta-learning scheme that tries to build a good learning algorithm based on a group of weak classifiers. In boosting, weighting is used to give more weight to more successful models. It can be applied to any classification learning algorithm. By weighting the instances, the learning algorithm can be forced to concentrate on a particular set of instances with more weight. Such instances are important because there is a greater incentive to classify them correctly (12).

2.3.11 Decision Stump

Decision Stump is a weak learner consisting of one-level binary decision tree. It is usually used in conjunction with a boosting algorithm. It implements regression based on the mean-squared error or classification based on the entropy.

2.3.12 Multiboost AB

Multiboosting is an extension to the Adaboost technique for forming decision committees. It can be viewed as a combination of Adaboost and wagging (a variant of bagging) techniques, combining the high bias of Adaboost technique and variance reduction property of wagging. This technique produces lower error than Adaboost or wagging. C4.5 is used as the base learner (14).

2.3.13 JRip

JRip implements a propositional rule learner called the ripper algorithm, an acronym for repeated incremental pruning to produce error reduction including heuristic global optimization of the rule set. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced-error pruning.

2.3.14 Random Committee

Random Committee builds an ensemble of randomized base classifiers and averages their predictions. Each base classifier is based on the same data but uses a different random number seed. This only makes sense if the base classifier is randomized, otherwise all classifiers would be the same (12).

2.3.15 Logistic Regression

This algorithm implements a multinomial logistic regression model with a ridge estimator. Logistic regression is a model used for prediction of the probability of occurrence of an event by fitting data into a logistic curve. There are some modifications in the implementation compared to the original logistic regression which does not deal with instance weights. The algorithm is modified a little bit to handle the instance weights (15). Ridge regression is a good method for obtaining more stable parameter estimates for the logistic regression model.

2.4 Survival Analysis

Survival analysis is a branch of statistics dealing with the death in biological organisms and failure in mechanical systems. Survival analysis examines and models the time it takes for events to occur. In our context, death from diseases can be considered as an event in the survival analysis. Survival models can be imagined to consist of two parts: the underlying hazard function describes how hazard (risk) changes over time and the effect parameters describe how hazard relates to other factors such as the choice of treatment, as in a medical scenario. When applied in the area of bioinformatics, survival analysis attempts to answer questions such as: what fraction of a population is expected to survive past a certain time? Of those that survive, at what rate will they die? Can multiple causes of death be taken into account? How do particular circumstances or characteristics increase or decrease the odds of survival? [15]

2.4.1 Cox proportional hazards model

Proportional hazards models are a sub-class of survival models in statistics, based on the assumption that effect parameters multiply hazard. For example, if taking drug X halves the hazard at time 0, it also halves the hazard at time 1, or at time t for any value of t. The effect parameters estimated by any proportional hazards model can be reported as hazard ratios. Sir David Cox observed that if the proportional hazards assumption holds (or, is assumed to hold) then it is possible to estimate the effect parameter(s) without any consideration of the hazard function[16]. This approach to survival data is called application of the Cox proportional hazards model. It is a broadly applicable and the most widely used method of survival analysis for exploring the relationship between the survival of a patient and several explanatory variables (16).

$$h_i(t) = h_0(t)\text{exp}\ (\beta_1 x_{i1} + \beta_2 x_{ik} + \ldots\ldots + \beta_k x_{ik}) \qquad \textit{(Equation 11)}$$

The baseline hazard function is given as α (t) = log $h_0(t)$

The above equation represents a semi-parametric model as the baseline hazard model. It can take any form where *i* represents the subscript for observation, *x* represents the covariates, constant *α* represents the log-baseline hazard.

[15] http://en.wikipedia.org/wiki/Survival_analysis

[16] http://en.wikipedia.org/wiki/Cox_regression

$$\text{or} \quad log\ h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{ik} + \cdots + \beta_k x_{ik} \qquad \textit{(Equation 12)}$$

Consider two observations *i* and *i'* that differ in their *x*-values, with the corresponding linear predictors as follows:

$$n_i = \beta_1 x_{i1} + \beta_2 x_{ik} + \cdots + \beta_k x_{ik} \qquad \textit{(Equation 13)}$$

$$n_{i\prime} = \beta_1 x_{i\prime 1} + \beta_2 x_{i\prime k} + \cdots + \beta_k x_{i\prime k} \qquad \textit{(Equation 14)}$$

The hazard ratio for these two observations is as follows:

$$\frac{hi(t)}{hi(t)} = \frac{h_0(t)e^{n_i}}{h_0(t)e^{n_{i}\prime}} \qquad \textit{(Equation 15)}$$

$$= \frac{e^{n_i}}{e^{n_{i}\prime}} \qquad \textit{(Equation 16)}$$

Given the survival times, status (alive or dead) and one or more covariates, Cox proportional hazards model produces a baseline survival curve, covariate coefficient estimates and their standard errors, risk ratios, 95% confidence intervals, and significance levels. A positive regression coefficient implies that the hazard is higher and thus the prognosis is worse for higher values. Conversely, a negative regression coefficient implies a better prognosis for patients with higher values of that variable.

2.4.2 Kaplan-Meier curves

Survival curves plot percentage of survival as a function of time. The Kaplan-Meier method is one of the techniques used for plotting survival curves. It is used to find out the proportion of the patients living for a certain amount of time after the treatment. The advantage of the Kaplan-Meier curve is that it takes into account, the "censored" data. A plot of the Kaplan-Meier estimate of the survival function is a series of horizontal steps of declining magnitude. In the Kaplan-Meier method, survival is recalculated every time a patient dies (17).

To calculate the fraction of patients who survived in a particular interval of time, divide the number alive at the end of the interval by the number alive at the beginning of the interval (excluding any censored patient in that interval from both the numerator and the denominator). This method automatically accounts for censored patients, as both the numerator and denominator are reduced for the interval when a patient is censored (18).

2.4.2.1 Kaplan-Meier estimator

Consider that a cohort has *n* individuals and t_1 , t_2, t_3........denote the actual times of death of the *n* individuals and d_1, d_2, d_3 …… denote the number of deaths that occur at each of these times. Let n_1, n_2 ,n_3……be the corresponding number of patients remaining in the cohort.

$$S(t) = \prod_{i=1}^{j}(1 - \frac{d_i}{n_i}) \qquad \textit{(Equation 17)}$$

The above equation represents the Kaplan-Meier estimator of the survival function S(t).

2.4.2.2 Interpretation of Kaplan-Meier Curves

- The Y- axis represents the estimated probability of survival.
- Precision of estimates depends on the number of observations, so the estimates on the left-hand side are more precise than the ones on the right-hand side. This is due to the less number of deaths and censored cases.
- But if a patient dies during the trial, then the survival curve reflects the patient's death at the appropriate time interval with a step down.
- The curve takes a step down every time a patient dies.
- The small blips or vertical tick-marks on the curve indicate when (time) the patient has been censored.
- Probability of surviving to any point is estimated from cumulative probability of surviving in each of the preceding time intervals (calculated as the product of preceding probabilities).

- There is another effect of censoring on the curve. As the patients are censored it reduces the number of patients contributing to the curve, so each death occurring after censoring represents a higher proportion of the remaining patients, and so every step down afterwards will be a bit larger than it would have been.

2.4.3 Log-rank test

Log-rank test is used to compare the survival of two groups of patients. Consider a survival plot showing two survival curves, one for low-risk group and the other for high-risk group. Looking at the curves, one can arrive at a conclusion that the low-risk group differs from the high-risk group (or vice versa) at an arbitrary time point, but nothing can be said about the two groups looking at the total survival time span. So we use the log-rank test which tells us whether the two groups differ significantly or not. The log-rank test is used to test the null hypothesis that there is no difference between the populations in the probability of an event (e.g. death) at any time point. A value of $p < 0.05$ indicates that the difference between the two groups is statistically significant. The log-rank test assumes that censoring is unrelated to the prognosis, and the survival probabilities are the same for subjects irrespective of the times when they were enrolled in the study. It is only a test of significance and it cannot provide an estimate of the difference between the groups or a confidence interval (19).

2.4.4 Time-dependent ROC curves

ROC curves display sensitivity and specificity of a continuous diagnostic marker for a binary disease variable. Time-dependent ROC curves take the disease outcome into account and vary as a function of time. In our study the binary disease variable R(t) = 1, if the patient had recurrence prior to time t, otherwise R(t) = 0. For a diagnostic marker M, both sensitivity and specificity are defined as a function of time t, as follows:

$$\textit{Sensitivity}(c,t)=P\{M>c|R(t)=1\} \qquad \textit{(Equation 18)}$$

$$\textit{Specificity}(c,t)=P\{M\leq c|R(t)=0\} \qquad \textit{(Equation 19)}$$

A time-dependent ROC curve is a plot of *1 – specificity(c, t)* versus *sensitivity(c, t)* for all possible values of threshold *c*. Sensitivity and specificity can be used to quantify the diagnostic ability of the test. Sensitivity is the probability that the test is positive, given that the person has the disease. Specificity is the probability that the test is negative, given that the person does not have the disease (20). The higher the ROC curve, the better is its capacity for discriminating diseased from non diseased subjects. ROC curves can also be used for comparing the discriminatory capacity of different diagnostic markers. In our study, the disease status changes with time. Some patients die as time progresses due to the disease or recurrence. So, we use time-dependent ROC curves instead of the classical ROC curves. There are different estimators for the ROC curves. We use the Kaplan-Meier based simple estimator in our ROC analysis.

2.5 Correlation coefficient

Correlation coefficient[17] indicates the strength of the relationship between two random variables. The correlation coefficient $\rho_{X,Y}$ between two random variables *X* and *Y* with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X\sigma_Y} = \frac{E((X-\mu X)(Y-\mu Y))}{\sigma_X\sigma_Y} \qquad \textit{(Equation 20)}$$

If we have a series of n measurements of X and Y written as x_i and y_i where $i = 1, 2, ..., n$, then the Pearson product-moment correlation coefficient can be used to estimate the correlation of X and Y. The Pearson correlation coefficient is given by the formula mentioned below.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}} \qquad \textit{(Equation 21)}$$

Correlation analysis is frequently used in microarray data analysis to measure the association between the variables. In our research, correlation analysis is used in validating cDNA microarray data by finding the

[17] http://en.wikipedia.org/wiki/Correlation

correlation between the gene copy number and RNA expression. This might be useful in understanding the genomic and proteomic level alterations in patients.

2.6 Related studies

Machine learning techniques and algorithms have been applied on microarray data from long time for tumor classification, prognosis prediction, and drug response prediction. This section describes some of the studies in the areas of tumor classification and prognosis prediction which are relevant to our research.

The study by Kwon et al (21) identified the genes involved in the carcinogenesis and progression of colorectal cancer by analyzing the gene-expression profiles of colorectal cancer cells using cDNA microarray. The samples and genes were classified by using a two way clustering analysis which identified genes that were differentially expressed in the cancerous and noncancerous tissues. Genes associated with lymph node metastasis were identified by using the *k*-nearest neighbors method. A 60-gene predictor correctly classified 10 of 12 patients (83.3%) as having colorectal cancer with lymph node metastasis versus those without metastasis.

The study by Koehler et al (22) created gene expression profiles from 25 colorectal carcinomas, corresponding normal colonic mucosa, and 14 liver metastases using cDNA arrays containing 1176 cancer related genes. Hierarchical clustering clearly distinguished carcinomas from non-cancerous tissues, separated tumors into high-stage and low-stage groups, and correlated with the histopathological classification in 87.0% of the cases. Statistical analysis (Mann–Whitney *U* test) revealed 40 tumor-specific genes which allowed identification of malignant tissue samples by clustering analysis. A specific expression signature in matching metastases was not found, but a set of 23 genes with statistically significant expression patterns ($p < 0.001$) in high and low stage tumors were identified.

The study by Croner et al (23) calculated the prediction rates for lymphatic metastasis using conventional clinicopathological parameters, gene expression data, and a combination of both. Prediction error, specificity, and sensitivity were analyzed using six different statistical classifiers. Analysis of conventional parameters produced a positive prediction rate that ranged between 53% and 61%. Microarray prediction rates were between 62.0% and 67.0% for lymphatic metastasis. It was concluded that the prediction of lymphatic metastasis can be improved by gene expression profiling of the primary tumor biopsy alone, or in combination with conventional parameters.

The study by Barrier et al (24) aimed at building a prognosis predictor that could be used for both stage II and stage III colon cancer patients to identify patients at high-risk of recurrence. The *k*-nearest neighbor classifier was used as a predictor. The main parameters of this classifier, the number of informative genes and the nearest neighbors *k* were chosen using cross validation. For both types of predictors (non-neoplastic mucosa and tumor based), 150 different pairs of parameters were considered and the performance of the corresponding predictors was assessed using six-fold cross-validation. Based on the results of cross validation, a 30-gene tumor based predictor and a 70-gene non-neoplastic mucosa based predictor were built on the whole set of patients. As a second set of independent samples was not available, a double cross-validation design was used, with an 'inner level' six-fold cross-validation for parameter selection and an 'outer level' three-fold cross-validation for performance assessment of the selected predictor. The estimated accuracy of the 30-gene tumor based predictor was 78.0% and that of the 70-gene non-neoplastic mucosa based predictor was 83.0%.

The study by Barrier et al (25) focused on identifying a subgroup of patients at high-risk of recurrence who were more likely to benefit from adjuvant chemotherapy based on non-neoplastic mucosa microarray gene expression measures of 24 patients (10 with a metachronous metastasis, 14 with no recurrence), for stage II colon cancer patients. The gene expression data of 24 patients was profiled using the Affymetrix HGU133A Gene Chip. A 70-gene prognosis predictor was identified, by selecting the 70 most differentially expressed genes (the number of genes to include was set to 70 based on the previous results) (24). A prognosis predictor was constructed by applying linear discriminant analysis on the 70-

gene set with a mean prognosis prediction accuracy of 81.8%, a sensitivity of 73.0%, and a specificity of 87.1% on the validation set.

The study by Bandres et al (5) aimed at identifying patients at high-risk of recurrence within the group of Duke's stage B patients. Tumor gene expression profiles from patients with Duke's B colorectal cancer were analyzed by high density oligonucleotide microarrays. The results showed that a subset of 48 genes were differentially expressed with an associated probability $P < 0.001$ in the t-test[18]. Another 11 genes, separating both the groups were identified using the Fisher criterion. Finally, 8 genes common in both the subsets were selected. The 8-gene signature was associated with relapse in Duke's stage B colon cancer patients, and it was able to discriminate between relapsed and non-relapsed patients. Furthermore, the differential expression of five genes *(CHD2, RPS5, ZNF148, BRI3 and MGC23401)* in colon cancer progression was confirmed by real-time PCR in an independent set of patients of Duke's B and C stages.

2.7 Open Problems

Microarray gene expression data is high dimensional, typically containing tens of thousands of features and a small sample size. Many of the genes contain irrelevant information which is not necessary for classification of the disease or phenotypes. Inclusion of these irrelevant genes increases the dimensionality of the dataset, introduces noise, and increases the computation time due to the complex search space. The data we analyzed in this study consisted of 73 observations of the expression levels of each of the 10,220 genes. Due to the very few observations and many features, innovative feature selection schemes need to be developed. Most of the studies described in the previous section explored the microarray gene expression data by using a single feature selection technique. Usually, a single feature selection technique is not enough to identify powerful gene signatures predicting the disease outcome given the high dimensional nature of the microarray data. Hence, we developed a combinatorial scheme to identify gene signatures. In the first step, we use random forests for variable selection. In the

[18] http://www.socialresearchmethods.net/kb/stat_t.php

second step, different attribute selection schemes in Weka such as CfsSubset, GainRatio, InfoGain and Relief were tested to reduce the feature set size by dropping some lower ranked features. It was found that the combination of random forests and InfoGain yielded the best results. This combinatorial feature selection scheme using random forests and InfoGain yields an optimal feature subspace which differentiates well between the classes in our study.

2.8 Summary

This chapter described the variable selection methods using random forests, various classification algorithms used in the study, the techniques used in survival analysis, the correlation coefficient analysis, related studies and open problems. The following chapters describe in detail how these techniques were applied on Ried et al colon cancer data to identify gene signatures, and results on independent colon cancer and rectal cancer datasets.

Chapter 3

Lymph node metastasis prediction model

3.1 Introduction

Accurately predicting the lymph node status or the stage of a cancer patient helps in selecting the optimal treatment. Staging is an important prognostic factor in determining the treatment options. The 5-year survival rate[19] in colon cancer patients with stage II tumors is ~78% and stage III tumors is ~64%. When a patient has been diagnosed with cancer, various clinical parameters are used to assess the risk of metastasis and death. In spite of the numerous advances in this area, tumor stage cannot be accurately determined by morphological assessment. With the advent of cDNA microarray technology, it is possible to measure the expression levels of thousands of genes simultaneously. Molecular markers identified based on the cDNA microarray gene expression data, have the ability to detect differences between the tumors at the molecular level (38). They offer improved prognostic accuracy when compared to the traditional methods. Patients at high-risk of metastasis can be identified and treated aggressively, while sparing other patients from the harmful effects of the invasive treatment. This chapter focuses on the identification and validation of the 9-gene lymph node status signature based on the microarray gene expression data in colon cancer patients.

Microarray gene expression data is highly correlated and many of the genes contain irrelevant information which is not necessary for classification of the disease or phenotypes. So, t-test[20] was done on genes with more than 5 missing values to evaluate the difference in proportions of missing values in node positive versus negative groups. Genes passing the t-test along with all genes having less than 5 missing values, a total of 10,220 genes were included for further analysis. We used a novel technique to identify biomarkers predicting the cancer stage. In the first step, variable selection using random forests was done

[19] http://www.webmd.com/colorectal-cancer/guide/treatment-stage

[20] http://www.socialresearchmethods.net/kb/stat_t.php

which identified a set of 14 genes. In the next step, the genes that did not have differential expression in lymph node negative versus positive tumors were discarded, leading to the identification of the 9-gene signature.

The discriminatory power of the 9-gene signature was evaluated by time-dependent ROC. The area under curve (AUC) was 0.85 and 0.86 for relapse-free survival (RFS) and overall survival (OS), respectively. The 9-gene signature generated significant patient stratification into low-risk and high-risk groups with distinct (*p=1e-04, n=73,* log-rank tests) and (*p=0.043, n=73,* log-rank tests) relapse-free survival (RFS) and overall survival (OS), respectively.

The remainder of this chapter is organized as follows. Section 3.2 introduces the data sets used for validation in the experiments. Section 3.3 introduces our study design and experiments in detail. Section 3.4 describes the validation results on multiple colon cancer datasets. Section 3.5 describes the correlation analysis, and Section 3.6 summarizes this chapter.

3.2 Description of the data sets

Ried et al PMID 17210682: The colon cancer microarray data from Ried et al contained 22,464 genes and 73 patient samples, all of them treated for primary adenocarcinomas of the colon. Of these 33 tumor samples were stage II (lymph node negative) and 40 tumor samples were stage III (lymph node positive). The relapse-free survival (RFS), overall survival (OS), and recurrence information was available for each of the patients in this dataset (2).

Koinuma et al PMID 16247484: The data used in this study was obtained from 10 specimens from each group (MSI $^{-}$ and MSI $^{+}$) subjected to gene expression profiling with microarrays. Affymetrix Gene Chip Human Genome U133 Array Set HG-U133 A and B was used in this analysis. The clinical information consisted of the Duke's stage for each of the patients (27).

Duke's Stage A   lymph node negative (class b)

Duke's Stage B lymph node negative (class b)

Duke's Stage C lymph node positive (class a)

Duke's Stage D these samples were not considered for leave-one-out cross validation.

Barrier et al PMID 16091735: This colon cancer data set consisted of 18 patient samples and 22,283 genes. The recurrence status (yes/no) was the available clinical information. Nine of the 18 patients developed a distant metastasis in the follow-up and the other nine patients remained disease-free for at least 5 years. All the patients were operated on for colonic adenocarcinomas. Ten patients had no lymph node metastasis (stage II) and did not receive any chemotherapy. The other eight patients had lymph node metastasis (stage III) and received 6-month adjuvant chemotherapy with fluorouracil (FU) and levamisole (24).

Barrier et al PMID 16966692: This colon cancer data set consisted of 50 patient samples and 22,283 genes. The recurrence status (yes/no) within 5 years was available in clinical data. Twenty-five patients developed a distant metastasis in the follow-up and the other 25 patients remained disease-free for at least 5 years. All the fifty patients were operated on for a stage II colon adenocarcinoma and none of the patients received any adjuvant chemotherapy (28).

Barrier et al PMID 17043639: This colon cancer data set consisted of 24 patients and 22,283 genes. The recurrence status (yes/no) within 5 years was also given. Ten patients developed a liver metastasis after surgery and the other 14 patients remained disease-free for at least 5 years. All the twenty-four patients were operated on for stage II colon adenocarcinomas and none of these 24 patients received any adjuvant chemotherapy (25).

NCI-60 data: The NCI-60[21] data contains a panel of 60 diverse human cancer cell lines used by the Developmental Therapeutics Program of the U.S. National Cancer Institute to screen >100,000 compounds and natural products. The RNA expression data for the cell lines is available under the

[21] http://discover.nci.nih.gov/cellminer/loadDownload.do

Affymetrix HG-U133A and HG-U133B chips. The drug activity data of 5-FU (fluorouracil) on all the 60 cell lines is available for download online[22].

Defining drug sensitivity and resistance: The drug activity profiles of 118 cancer agents including 5-FU are available online. 5-FU is the drug frequently used in colon cancer treatment. The recorded drug activities ($\log_{10}$ GI_{50}) were available for the 60 human cancer cell lines. Specifically, for each drug, $\log_{10}$ (GI_{50}) values were normalized across the 60 cell lines. Cell lines with $\log_{10}$ (GI_{50}) at least 0.5 SD above the mean were defined as resistant to the drug. Those with $\log_{10}$ (GI_{50}) at least 0.5 SD below the mean were defined as sensitive to the drug. The remaining cell lines with $\log_{10}$ (GI_{50}) within 0.5 SD were defined as intermediate in the range of drug responses (41). Specifically, 17 cell lines were sensitive, 26 cell lines were intermediate, and the other 17 cell lines were resistant to the drug Fluorouracil (5-FU).

[22] http://discover.nci.nih.gov/nature2000/data/selected_data/dataviewer.jsp?baseFileName=a_matrix118&nsc=2&dataStart=3

3.3 Study Design for the 9-gene signature

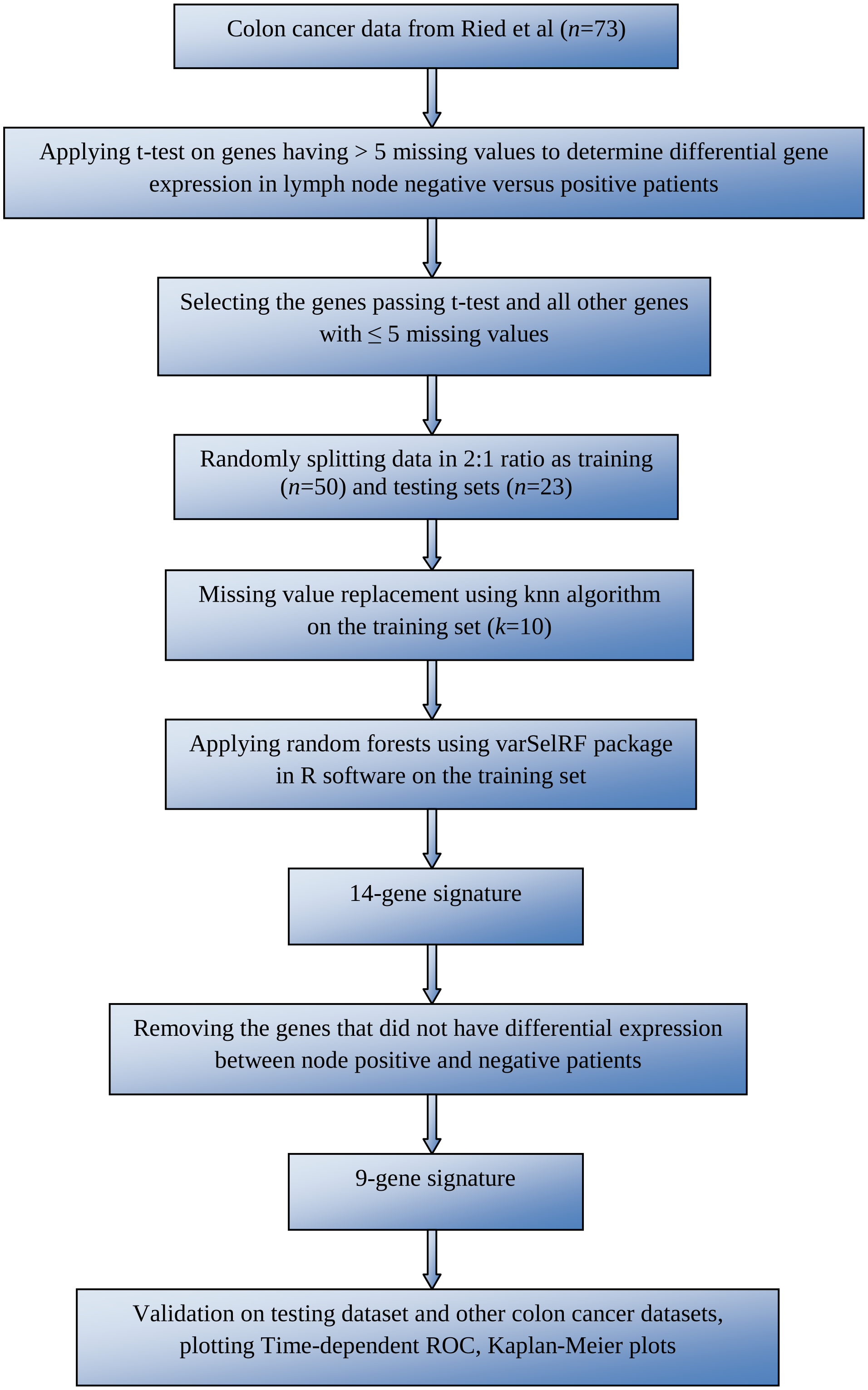
Colon cancer data from Ried et al (*n*=73)
Applying t-test on genes having > 5 missing values to determine differential gene expression in lymph node negative versus positive patients
Selecting the genes passing t-test and all other genes with ≤ 5 missing values
Randomly splitting data in 2:1 ratio as training (*n*=50) and testing sets (*n*=23)
Missing value replacement using knn algorithm on the training set (*k*=10)
Applying random forests using varSelRF package in R software on the training set
14-gene signature
Removing the genes that did not have differential expression between node positive and negative patients
9-gene signature
Validation on testing dataset and other colon cancer datasets, plotting Time-dependent ROC, Kaplan-Meier plots

Figure 3.1 Block diagram of the study for 9-gene signature.

3.3.1 Experimental procedure

Data Source The colon cancer microarray data from Ried et al. (2) contained 22,464 genes and 73 patient samples, all of them treated for primary adenocarcinomas of the colon. Of these 33 tumor samples were stage II (lymph node negative) and 40 tumor samples were stage III (lymph node positive).

Log Ratio Every spot on the microarray provides two intensity values each of them associated with a specific channel. Dividing one intensity by the other gives the expression ratio. We use log ratios as they are lot easier to work with than the regular ratios. The log ratio (532/635) was considered for this analysis. It is the log (base 2) transformation of the ratio of medians at wavelengths of 532nm and 635 nm.

Data Preprocessing - t test We investigated whether the observed difference between the two groups (node positive versus negative) represents a real difference in the total study population from which the sample was drawn, or whether it just occurred by chance (due to sampling variation), by using t-test. The number of missing values for each gene was found and t-test was done on genes with more than 5 missing values to evaluate the difference in the proportions of missing values in node positive versus negative groups. The genes passing the t-test ($p < 0.05$, two-sided) were included along with all genes having less than 5 missing values for further processing. A total of 10,220 genes satisfied this condition.

Training dataset The data obtained in the above step was randomly split in 2:1 ratio as training set and testing set. The expression data of the 10,220 genes and 50 patients constituted the training set.

Missing value replacement The training dataset contained missing values. They were replaced using the EMV[23] package in R software with k=10. This technique estimates the missing values based on the k-nearest neighbors algorithm. This algorithm selects the k nearest rows that do not contain any missing values to the one containing at least one missing value based on the Euclidian distance. Then the missing values are replaced by the average of the neighbors.

[23] http://cran.r-project.org/web/packages/EMV/index.html

Biomarker identification VarSelRF[24] package in R was used in a series of steps on the training dataset to find the important features. Lymph node status was used as the class variable. In the first step, a forest with N trees was built and the features were ranked according to the importance of the variables. In the second step, 20% of the variables that were least important were removed and a new forest was constructed with K trees. This step was repeated till there were two genes left. In the experiment, a value of N = 2000 and K =1000 were considered, because a large number of trees in the initial forests is likely to produce stable importance measures (23). After fitting all forests, the OOB error rates from all the fitted random forests were examined and a set of 15 genes leading to the smallest error rate were selected. There was a control gene in the identified 15 genes which was discarded leaving 14 genes. Table 3.1 shows the 14 genes.

Table 3.1 The 14-gene lymph node status signature.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
IFRG28-28kD interferon responsive pro	H200004627
PDCD5-programmed cell death 5	H200007687
PLXNB2-plexin B2, mRNA	H200000861
HIST1H3I-histone 1,H3i,m	H200013045
DC50-hypothetical protein DC50	H200019106
SR140-U2-associated SR140 protein	H200020644
FLJ11078-hypothetical protein FLJ1107	H200016227
MGC16044-hypothetical protein MGC1604	H200020589
RNF6-ring finger protein (C3H2C3 type)	H200004174
POU6F2-POU domain, class 6,transcript	H200015474
LAMB1-laminin,beta 1(LAMB1), mRNA	H200006892
ITGB1-integrin,beta1 (fibronectin)	H200021334
HIST1H2BO-histone 1, H2bo HIST1H2BO	H200013772

[24] http://cran.r-project.org/web/packages/varSelRF/index.html

3.3.2 Differentially expressed genes

Differentially expressed genes or discriminator genes are the genes with significantly different expression in the two user defined groups or between samples obtained under different conditions in a gene expression experiment. These gene signatures or disease associated markers are relevant to biological processes. To find the differentially expressed genes, the mean expression values for each of the 14 genes were calculated for lymph node negative and positive tumor groups separately. If the gene had a higher value in node positive versus negative samples it was over expressed and vice versa. Table 3.2 shows the over expressed and under expressed genes, and *p*-values for each gene obtained using the z-test[25]. The genes that did not have differential expression among lymph node negative and positive patients, namely, *PDCD5, HIST1H3I, SR140, LAMB1, and HIST1H2BO* in the 14-gene signature were removed. Table 3.3 shows the remaining 9 genes.

Table 3.2 Over expressed and under expressed genes in the 14-gene signature between lymph node positive and negative patients.

GENE NAME	Category in lymph node positive group	*p*-value	Significance
SNRPD3-small nuclear ribonucleoprotein	Under expressed	0.041134	Yes
IFRG28-28kD interferon responsive pro	Under expressed	0.044768	Yes
PDCD5-programmed cell death 5	Under expressed	0.062525	No
PLXNB2-plexin B2, mRNA	Under expressed	0.038442	Yes
HIST1H3I-histone 1,H3i,m	Under expressed	0.103315	No
DC50-hypothetical protein DC50	Under expressed	0.003831	Yes
SR140-U2-associated SR140 protein	Over expressed	0.152485	No
FLJ11078-hypothetical protein FLJ1107	Under expressed	0.002202	Yes
MGC16044-hypothetical protein MGC1604	Over expressed	0.006503	Yes
RNF6-ring finger protein (C3H2C3 type)	Over expressed	0.008206	Yes
POU6F2-POU domain, class 6,transcript	Over expressed	0.002299	Yes
LAMB1-laminin,beta 1(LAMB1), mRNA	Under expressed	0.166319	No
ITGB1-integrin,beta1 (fibronectin)	Over expressed	0.012588	Yes
HIST1H2BO-histone 1, H2bo HIST1H2BO	Under expressed	0.053911	No

[25] http://en.wikipedia.org/wiki/Z-test

Table 3.3 The 9-gene signature for predicting lymph node metastasis.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
IFRG28-28kD interferon responsive pro	H200004627
PLXNB2-plexin B2, mRNA	H200000861
DC50-hypothetical protein DC50	H200019106
FLJ11078-hypothetical protein FLJ1107	H200016227
MGC16044-hypothetical protein MGC1604	H200020589
RNF6-ring finger protein (C3H2C3 type)	H200004174
POU6F2-POU domain, class 6,transcript	H200015474
ITGB1-integrin,beta1 (fibronectin)	H200021334

3.4 Results

3.4.1 Validation of the 9-gene signature on testing data (*n=23*)

The original data was split in 2:1 ratio as training and testing datasets, respectively. The testing data consisted of 23 tumor samples. Eleven tumor samples were lymph node negative and the other 11 samples were lymph node positive. The data used for validation consisted of the expression of the 9-gene signature in the 23 patient samples. Weka software was used for validation and lymph node status (negative/positive) was predicted. Different classification schemes including *J48*, Logistic regression, *KStar*, Threshold selector, and Multilayer perceptron were applied to this dataset to find the best scheme. Table 3.4 shows the comparison between *J48* and some of the classifiers used for validation on other datasets. *J48* classifier performed better than the other classifiers. It had a sensitivity of 75.00%, a specificity of 81.80%, and an overall accuracy of 78.26%. Table 3.5 shows the confusion matrix for *J48* classifier. The difference in overall accuracy between *J48* and other classifiers was not statistically significant due to the small sample size.

Table 3.4 Comparison of accuracies obtained from different classifiers for predicting lymph node status using the 9-gene signature. The improved overall accuracy of the prediction with the *J48* classifier compared with other methods was assessed by significance testing (*N* = 23).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
J48	**75.00**	**81.80**	**78.40**	**78.26**	
Logistic regression	66.70	54.50	60.60	60.86	<0.11
KStar	66.70	54.50	60.60	60.86	<0.11
Threshold selector	58.30	72.70	65.50	65.21	<0.17
Multilayer perceptron	66.70	45.50	56.10	56.52	<0.06

Table 3.5 Confusion matrix obtained from the *J48* classifier for predicting lymph node status using the 9-gene signature.

Actual/Predicted	**a (node negative)**	**b (node positive)**
a (node negative)	9	2
b (node positive)	3	9

3.4.2 Time-dependent ROC analyses on data from Ried et al (*n=73*)

To explore whether the 9-gene lymph node signature could predict patient disease-free survival and overall survival, the survival and status information along with the expression data of the 9 genes are used for getting the time-dependent ROC plots. The accuracy of 5-year relapse-free survival prediction using these 9 genes is 0.85 and 5-year overall survival prediction is 0.86, as represented by the AUC.

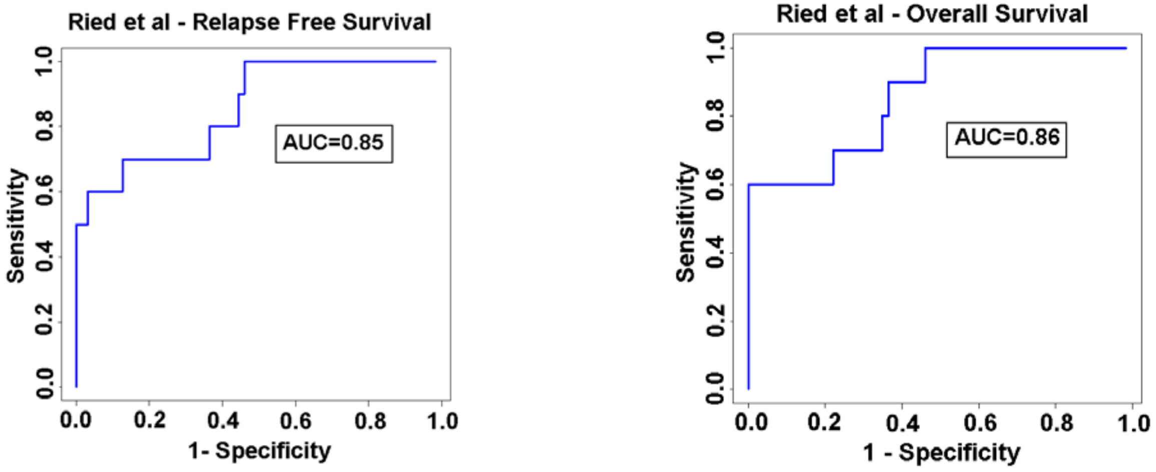

Figure 3.2 Time-dependent ROC plots on data from Ried et al (*n=73*) for relapse-free survival and overall survival using the 9-gene signature.

3.4.3 Kaplan-Meier analyses on data from Ried et al (*n=73*)

The Cox model based on the expression of the 9-gene signature was used to get recurrence risk scores for all the 73 patients. The choices for choosing a cut-off value for patient stratification are the peak value from histogram, mean risk score or median risk score. In this analysis, the peak value from histogram was chosen as cut-off as it resulted in best patient stratification. Cut-off values of 4.0 and 0.5 were chosen for relapse-free survival and overall survival, respectively. The *pamr* package in R was used to plot the Kaplan-Meier curves, for relapse-free survival and overall survival. The low-risk and high-risk groups had distinct relapse-free survival (*p = 1e-04, n=73,* log-rank tests) and overall survival (*p = 0.043, n=73*, log-rank tests).

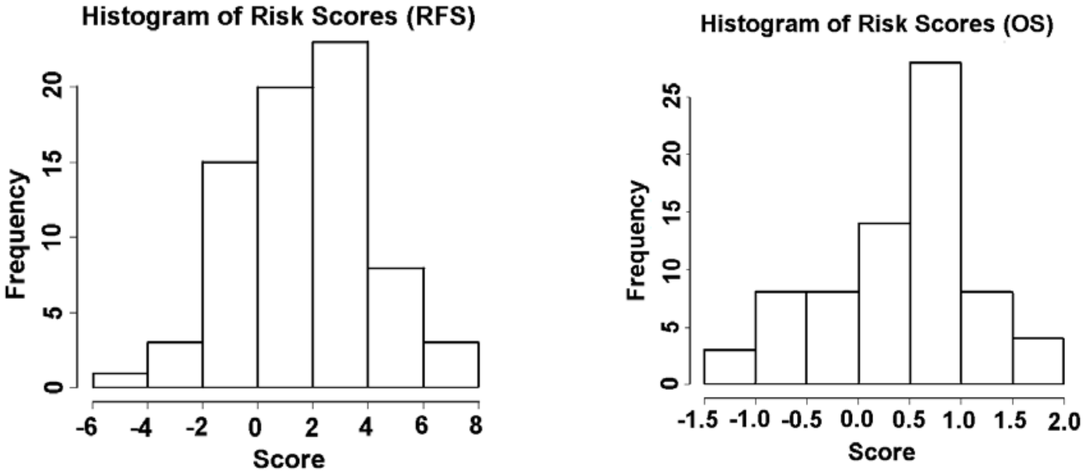

Figure 3.3 Histograms of risk scores obtained from Cox model for relapse-free survival and overall survival using the 9-gene signature.

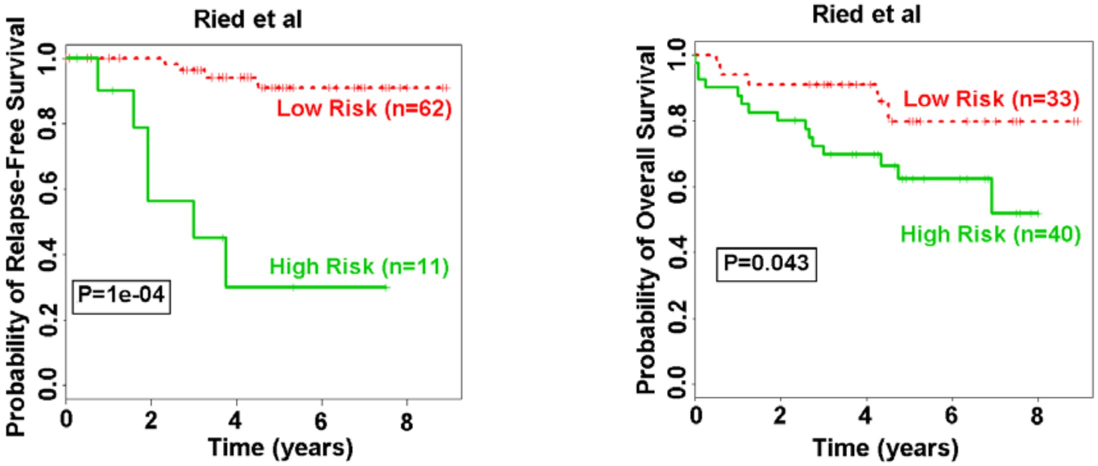

Figure 3.4 Kaplan-Meier plots on data from Ried et al (*n=73*) for relapse-free survival and overall survival using the 9-gene signature.

Out of the 73 patients in the colon cancer data from Ried et al, 26 patients remained relapse-free for more than 5 years and 10 patients experienced recurrence within 5 years after surgery. To test the performance of the identified 9-gene signature, the subgroups obtained for the above group of 36 patients from the Cox model were compared with their actual clinical outcomes. Table 3.6 shows the different parameters obtained from the Cox model, using the 9-gene signature for relapse-free survival and overall survival, respectively. Tables 3.7 and 3.8, show the comparison of predicted clinical outcome for patients with their actual follow-up information, for relapse-free survival and overall survival, respectively. The Cox model had a sensitivity of 60.0%, a specificity of 92.3%, and an overall accuracy of 83.3%, for predicting relapse-free survival. In predicting overall survival, it had a sensitivity of 75.0%, a specificity of 45.8%, and an overall accuracy of 59.0%.

Table 3.6 Different parameters obtained from the Cox model using the 9-gene signature for predicting relapse-free survival and overall survival.

Relapse-free survival						Overall survival				
Gene Symbol	**coef**	**exp (coef)**	**se (coef)**	**z-score**	***p*-value**	**coef**	**exp (coef)**	**se (coef)**	**z-score**	***p*-value**
SNRPD3	2.661	14.304	1.326	2.006	0.045	0.065	1.068	0.793	0.082	0.93
IFRG28	0.769	2.157	0.382	2.012	0.044	0.122	1.131	0.235	0.521	0.60
PLXNB2	-3.121	0.044	1.305	-2.391	0.017	-0.516	0.597	0.596	-0.866	0.39
DC50	-2.476	0.084	0.797	-3.107	0.001	-0.535	0.586	0.459	-1.166	0.24
FLJ11078	0.600	1.822	0.584	1.028	0.300	0.002	1.002	0.362	0.005	1.00
MGC16044	0.525	1.690	0.380	1.382	0.170	0.145	1.157	0.269	0.542	0.59
RNF6	-1.163	0.312	0.815	-1.426	0.150	0.145	0.574	0.465	-1.193	0.23
POU6F2	0.869	2.384	1.179	0.737	0.460	0.596	1.816	0.581	1.026	0.30
ITGB1	1.612	5.013	1.237	1.303	0.190	0.465	1.592	0.603	0.771	0.44

Table 3.7 Comparison of the sub groups predicted from the Cox model using the 9-gene signature with the actual subgroups for relapse-free survival.

	Recurrence	No recurrence	Sensitivity (%)	Specificity (%)	Overall accuracy (%)
Recurrence	6	4	60.0	92.3	83.3
No recurrence	2	24			

Table 3.8 Comparison of the sub groups predicted from the Cox model using the 9-gene signature with the actual subgroups for overall survival.

	Death	Alive	Sensitivity (%)	Specificity (%)	Overall accuracy (%)
Death	15	5	75.0	45.8	59.0
Alive	13	11			

The Cox model was used for stratifying all the 73 patients in Ried et al data, into low-risk and high-risk groups, based on the 9-gene signature. Out of the 73 patients, a total of 37 did not have recurrence with survival times less than 5 years. Twenty-nine patients had overall survival times less than 5 years without any event (death). The relapse outcome for the 37 patients and the overall survival outcome for the 29 patients is currently unknown. Table 3.9 shows the prospective prognostic predictions of these patients obtained from the Cox model for relapse-free survival and overall survival, respectively. The follow-up information for these patients is being collected. When it becomes available in the future, the predictions can be validated with it.

Table 3.9 Patient subgroups obtained from the Cox model for relapse-free survival and overall survival using the 9-gene signature.

Serial Number	Patient ID	Predicted group by Cox model (RFS)	Patient ID	Predicted group by Cox model (OS)
1	CC-P1	Low Risk	CC-P1	Low Risk
2	CC-P2	Low Risk	CC-P4	Low Risk
3	CC-P4	Low Risk	CC-P7	High Risk
4	CC-P7	Low Risk	CC-P8	Low Risk
5	CC-P8	Low Risk	CC-P9	Low Risk
6	CC-P10	High Risk	CC-P11	High Risk
7	CC-P13	Low Risk	CC-P13	Low Risk
8	CC-P18	Low Risk	CC-P16	Low Risk
9	CC-P20	Low Risk	CC-P18	Low Risk
10	CC-P21	Low Risk	CC-P20	Low Risk
11	CC-P22	Low Risk	CC-P21	Low Risk
12	CC-P23	Low Risk	CC-P22	Low Risk
13	CC-P25	Low Risk	CC-P25	Low Risk
14	CC-P28	Low Risk	CC-P28	Low Risk
15	CC-P29	High Risk	CC-P31	High Risk
16	CC-P31	Low Risk	CC-P35	High Risk
17	CC-P34	Low Risk	CC-P36	High Risk
18	CC-P35	Low Risk	CC-P37	High Risk
19	CC-P37	Low Risk	CC-P38	Low Risk
20	CC-P38	Low Risk	CC-P40	High Risk
21	CC-P40	High Risk	CC-P48	High Risk
22	CC-P42	Low Risk	CC-P50	High Risk
23	CC-P44	Low Risk	CC-P51	Low Risk
24	CC-P46	Low Risk	CC-P60	Low Risk
25	CC-P47	Low Risk	CC-P62	Low Risk
26	CC-P48	Low Risk	CC-P66	High Risk
27	CC-P50	Low Risk	CC-P71	High Risk
28	CC-P51	Low Risk	CC-P72	Low Risk
29	CC-P55	Low Risk	CC-P73	High Risk
30	CC-P56	Low Risk		
31	CC-P60	Low Risk		
32	CC-P62	Low Risk		
33	CC-P66	Low Risk		
34	CC-P68	Low Risk		
35	CC-P70	Low Risk		
36	CC-P71	Low Risk		
37	CC-P72	Low Risk		

3.4.4 External validation of the 9-gene signature on other colon cancer data

This part of the study sought to explore the extent to which the 9-gene signature could be used for prediction of lymph node status, recurrence, and drug response in publicly available independent datasets. More than 50 classifiers available in Weka software were tested using a leave-one-out cross validation technique on each of the independent datasets to find a suitable classification scheme for validation. Due to the different number of attributes (matching genes), sample sizes and prediction variables one specific scheme could not be used for validation on all the datasets. Different classifiers had to be employed on the validation datasets to get fair prediction accuracy. As far as possible the same set of classifiers were presented in the comparison tables of validation datasets to provide a fair evaluation of the performance. The exact same set of classifiers could not be compared over all the validation datasets due to poor performances of classifiers on some datasets and good performances on other datasets. The following sections discuss the validation results and comparisons of various classifiers on the independent datasets in detail.

3.4.4.1 Predicting lymph node status by leave-one-out cross validation on data from Koinuma et al (*n=17*) PMID 16247484

The data from Koinuma et al (Affymetrix HG U133 A platform) consisted of 20 patient samples of which 3 patients were Duke's stage D. The Duke's stage D patients were not considered for validation. The search for matching genes was done using the Affymetrix ids. There were 7 matching genes (Table 3.10). The data used for validation consisted of the expression of these 7 genes in the 17 patients. Weka software was used for cross validation and the lymph node status (positive/negative) was predicted. Different classification schemes including Naïve Bayes, LWL, JRip, Bagging, and *KStar* were applied to this dataset to find the best scheme. Table 3.11 shows the comparison between Naïve Bayes and some of the classifiers used for validation on other datasets. Naive Bayes classifier performed better than the other classifiers. It had a sensitivity of 100.00%, a specificity of 75.00%, and an overall accuracy of 82.35%.

Table 3.12 shows the confusion matrix for Naïve Bayes classifier. The difference in overall accuracy between Naïve Bayes and other classifiers was not statistically significant due to the small sample size.

Table 3.10 Matching genes in Koinuma et al data.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
IFRG28-28kD interferon responsive pro	H200004627
PLXNB2-plexin B2, mRNA	H200000861
FLJ11078-hypothetical protein FLJ1107	H200016227
RNF6-ring finger protein (C3H2C3 type)	H200004174
POU6F2-POU domain, class 6,transcript	H200015474
ITGB1-integrin,beta1 (fibronectin)	H200021334

Table 3.11 Comparison of accuracies obtained from different classifiers for predicting lymph node status using the 9-gene signature. The improved overall accuracy of the prediction with the Naïve Bayes classifier compared with other methods was assessed by significance testing (*N* = 17).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Naïve Bayes	**100.00**	**75.00**	**87.50**	**82.35**	
LWL	71.40	70.00	70.70	70.58	< 0.21
JRip	57.10	70.00	63.55	64.70	< 0.13
Bagging	57.10	70.00	63.55	64.70	< 0.13
KStar	42.90	70.00	56.45	58.82	< 0.07

Table 3.12 Confusion matrix obtained from the Naïve Bayes classifier for predicting lymph node status using the 9-gene signature.

Actual/Predicted	**a (node positive)**	**b (node negative)**
a (node positive)	7	0
b (node negative)	3	7

3.4.4.2 Predicting recurrence by leave-one-out cross validation on data from Barrier et al (*n=18*) PMID 16091735

The data from Barrier et al (PMID 16091735) consisted of 22,283 genes and 18 patient samples. The search for matching genes was done using the Affymetrix ids. There were 7 matching genes (Table 3.13). The data used for validation consisted of the expression of these 7 genes in the 18 patient samples. Weka software was used for validation and recurrence (yes/no) was predicted. Different classification schemes including Naïve Bayes, *KStar*, Multilayer perceptron, JRip, and Logistic regression were applied to this dataset to find the best scheme. Table 3.14 shows the comparison between Naïve Bayes and some of the classifiers used for validation on other datasets. Naive Bayes classifier performed better than the other classifiers. It had a sensitivity of 100.00%, a specificity of 88.90%, and an overall accuracy of 94.44%. Table 3.15 shows the confusion matrix for Naïve Bayes classifier. The difference in overall accuracy between Naïve Bayes and other classifiers was not statistically significant due to the small sample size.

Table 3.13 Matching genes in Barrier et al data.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
IFRG28-28kD interferon responsive pro	H200004627
PLXNB2-plexin B2, mRNA	H200000861
FLJ11078-hypothetical protein FLJ1107	H200016227
RNF6-ring finger protein (C3H2C3 type)	H200004174
POU6F2-POU domain, class 6,transcript	H200015474
ITGB1-integrin,beta1 (fibronectin)	H200021334

Table 3.14 Comparison of accuracies obtained from different classifiers for predicting recurrence using the 9-gene signature. The improved overall accuracy of the prediction with the Naïve Bayes classifier compared with other methods was assessed by significance testing (*N* = 18).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
NaiveBayes	**100.00**	**88.90**	**94.45**	**94.44**	
KStar	100.00	66.70	53.35	83.33	< 0.15
Multilayer perceptron	88.90	77.80	83.35	83.33	< 0.15
JRip	88.90	66.70	77.80	77.77	< 0.08
Logistic regression	88.90	66.70	77.80	77.77	< 0.08

Table 3.15 Confusion matrix obtained from the Naïve Bayes classifier for predicting recurrence using the 9-gene signature.

Actual/Predicted	**a (recurrence)**	**b (no recurrence)**
a (recurrence)	9	0
b (no recurrence)	1	8

3.4.4.3 Predicting recurrence by leave-one-out cross validation on data from Barrier et al (*n=50*) (PMID 16966692)

The data from Barrier et al (PMID 16966692) consisted of 22,283 genes and 50 patient samples. The search for matching genes was done using the Affymetrix ids. There were 7 matching genes (Table 3.16). The data used for validation consisted of the expression of these 7 genes in the 50 patient samples. Weka software was used for validation and recurrence (yes/no) was predicted. Different classification schemes including Decision stump, Naïve Bayes, IB1, Logistic regression, and Multilayer perceptron were applied to this dataset to find the best scheme. Table 3.17 shows the comparison between Decision stump and some of the classifiers used for validation on other datasets. Decision stump classifier performed better than the other classifiers. It had a sensitivity of 76.00%, a specificity of 88.00%, and an overall accuracy of 82.00%. Table 3.18 shows the confusion matrix for Decision stump classifier. The difference in overall accuracy between Decision stump and other classifiers was not statistically significant due to the small sample size.

Table 3.16 Matching genes in Barrier et al data.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
IFRG28-28kD interferon responsive pro	H200004627
PLXNB2-plexin B2, mRNA	H200000861
FLJ11078-hypothetical protein FLJ1107	H200016227
RNF6-ring finger protein (C3H2C3 type)	H200004174
POU6F2-POU domain, class 6,transcript	H200015474
ITGB1-integrin,beta1 (fibronectin)	H200021334

Table 3.17 Comparison of accuracies obtained from different classifiers for predicting recurrence using the 9-gene signature. The improved overall accuracy of the prediction with the Decision stump classifier compared with other methods was assessed by significance testing (*N* = 50).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity +Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Decision stump	**76.00**	**88.00**	**82.00**	**82.00**	
NaiveBayes	68.00	92.00	80.00	80.00	< 0.40
IB1	84.00	64.00	74.00	74.00	< 0.16
Logistic regression	68.00	72.00	70.00	70.00	< 0.08
Multilayer perceptron	68.00	72.00	70.00	70.00	< 0.08

Table 3.18 Confusion matrix obtained from the Decision stump classifier for predicting recurrence using the 9-gene signature.

Actual/Predicted	**a (no recurrence)**	**b (recurrence)**
a (no recurrence)	22	3
b (recurrence)	6	19

3.4.4.4 Predicting recurrence by leave-one-out cross validation on data from Barrier et al (*n=24*) (PMID 17043639)

The data from Barrier et al (PMID 17043639) consisted of 22,283 genes and 24 patient samples. The search for matching genes was done using the Affymetrix ids. There were 7 matching genes (Table 3.19). The data used for validation consisted of the expression of these 7 genes in the 24 patient samples. Weka software was used for validation and recurrence (yes/no) was predicted. Different classification schemes including Naïve Bayes, LWL, AD Tree, Random committee, and Multiboost AB were applied to this dataset to find the best scheme. Table 3.20 shows the comparison between Naïve Bayes and some of the classifiers used for validation on other datasets. Naive Bayes classifier performed better than the other classifiers. It had a sensitivity of 50.00%, a specificity of 100.00%, and an overall accuracy of 79.16%. Table 3.21 shows the confusion matrix for Naïve Bayes classifier. The difference in overall accuracy between Naïve Bayes and other classifiers was not statistically significant due to the small sample size.

Table 3.19 Matching genes in Barrier et al data.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
IFRG28-28kD interferon responsive pro	H200004627
PLXNB2-plexin B2, mRNA	H200000861
FLJ11078-hypothetical protein FLJ1107	H200016227
RNF6-ring finger protein (C3H2C3 type)	H200004174
POU6F2-POU domain, class 6,transcript	H200015474
ITGB1-integrin,beta1 (fibronectin)	H200021334

Table 3.20 Comparison of accuracies obtained from different classifiers for predicting recurrence using the 9-gene signature. The improved overall accuracy of the prediction with the Naïve Bayes classifier compared with other methods was assessed by significance testing (*N* = 24).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity +Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
NaiveBayes	**50.00**	**100.00**	**75.00**	**79.16**	
LWL	60.00	78.60	69.30	70.83	< 0.26
AD Tree	40.00	92.90	66.45	70.83	< 0.26
Random committee	40.00	85.70	62.85	66.66	< 0.17
Multiboost AB	50.00	71.40	60.70	62.50	< 0.11

Table 3.21 Confusion matrix obtained from the Naïve Bayes classifier for predicting recurrence using the 9-gene signature.

Actual/Predicted	**a (no recurrence)**	**b (recurrence)**
a (no recurrence)	14	0
b (recurrence)	5	5

3.4.4.5 Predicting the response of cell lines in NCI-60 (*n=34*) (U133A GCRMA) data by leave-one-out cross validation

This dataset[26] consisted of 21,225 genes and 60 cell lines (41). Our focus was on the sensitive and resistant cell lines, so cell lines with intermediate response were not considered for validation. A total of 34 cell lines (17 sensitive and the other 17 resistant to the drug 5-FU) were used in validation. The search for matching genes was done using the gene symbols. There were 5 matching genes (Table 3.22). The

[26] http://discover.nci.nih.gov/cellminer/loadDownload.do

data used for validation consisted of the expression of these 5 genes in the 34 cell lines. Weka software was used for validation and the response (sensitive/resistant) to the drug 5-FU (fluorouracil) was predicted. Different classification schemes including Threshold selector, *IB1*, Logistic regression, Random Tree, and Multilayer perceptron were applied to this dataset to find the best scheme. Table 3.23 shows the comparison between Threshold selector and some of the classifiers used for validation on other datasets. Threshold selector performed better than the other classifiers. It had a sensitivity of 94.10%, a specificity of 76.50%, and an overall accuracy of 85.29%. Table 3.24 shows the confusion matrix for Threshold selector classifier. The difference in overall accuracy between Threshold selector and *IB1* ($p < 0.01$), Logistic regression ($p < 0.01$), Random Tree ($p < 0.01$), Multilayer perceptron ($p < 0.01$) was statistically significant.

Table 3.22 Matching genes in NCI-60 U133A data.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
PLXNB2-plexin B2, mRNA	H200000861
RNF6-ring finger protein (C3H2C3 type)	H200004174
POU6F2-POU domain, class 6,transcript	H200015474
ITGB1-integrin,beta1 (fibronectin)	H200021334

Table 3.23 Comparison of accuracies obtained from different classifiers for predicting drug response using the 9-gene signature. The improved overall accuracy of the prediction with the Threshold selector classifier compared with other methods was assessed by significance testing (*N* = 34).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity +Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Threshold selector	**94.10**	**76.50**	**85.30**	**85.29**	
IB1	58.80	52.90	55.85	55.88	< 0.01
Logistic regression	52.90	47.10	50.00	50.00	< 0.01
Random Tree	47.10	52.90	50.00	50.00	< 0.01
Multilayer perceptron	35.30	41.20	38.25	38.23	< 0.01

Table 3.24 Confusion matrix obtained from the Threshold selector classifier for predicting drug response using the 9-gene signature.

Actual/Predicted	a (sensitive)	b (resistant)
a (sensitive)	16	1
b (resistant)	4	13

3.4.4.6 Predicting the response of cell lines in NCI-60 (*n=34*) (U133B GCRMA) data by leave-one-out cross validation

This dataset[27] consisted of 17910 genes and 60 cell lines (41). Our focus was on the sensitive and resistant cell lines, so cell lines with intermediate response were not considered for validation. A total of 34 cell lines (17 sensitive and the other 17 resistant to the drug 5-FU) were used in validation. The search for matching genes was done using the gene symbols. There was 1 matching gene (Table 3.25). The data used for validation consisted of the expression of this gene in the 34 cell lines. Weka software was used for validation and the response (sensitive/resistant) for the drug 5-FU (fluorouracil) was predicted. Different classification schemes including Threshold selector, Random Tree, Random committee, Decision stump, and AD Tree were applied to this dataset to find the best scheme. Table 3.26 shows the comparison between Threshold selector and some of the classifiers used for validation on other datasets. Threshold selector performed better than the other classifiers. It had a sensitivity of 82.40%, specificity of 76.50%, and an overall accuracy of 79.41%. Table 3.27 shows the confusion matrix for Threshold selector classifier. The difference in overall accuracy between Threshold selector and Random tree ($p < 0.02$), Random committee ($p < 0.02$), Decision stump ($p < 0.01$), AD Tree ($p < 0.01$) was statistically significant.

Table 3.25 Matching genes in NCI-60 U133B data.

GENE NAME	ID
POU6F2-POU domain, class 6,transcript	H200015474

[27] http://discover.nci.nih.gov/cellminer/loadDownload.do

Table 3.26 Comparison of accuracies obtained from different classifiers for predicting drug response using the 9-gene signature. The improved overall accuracy of the prediction with the Threshold selector classifier compared with other methods was assessed by significance testing (*N* = 34).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity +Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Threshold selector	**82.40**	**76.50**	**79.45**	**79.41**	
Random Tree	64.70	47.10	55.90	55.88	< 0.02
Random committee	64.70	47.10	55.90	55.88	< 0.02
Decision stump	94.10	11.80	52.95	52.94	< 0.01
AD Tree	52.90	47.10	50.00	50.00	< 0.01

Table 3.27 Confusion matrix obtained from the Threshold selector classifier for predicting drug response using the 9-gene signature.

Actual/Predicted	**a (sensitive)**	**b (resistant)**
a (sensitive)	14	3
b (resistant)	4	13

3.4.5 Summary of validation results of 9-gene signature

Table 3.28 shows the details of different validation datasets, predicted variables, classifiers used and different accuracies obtained using the 9-gene signature. For each dataset the classifier with the highest overall accuracy was reported.

Table 3.28 Summary of validation results of 9-gene signature on Ried et al data, independent colon cancer datasets and NCI-60 data.

Dataset	Classifier	Predicted variable	Sensitivity (%)	Specificity (%)	(Sensitivity + Specificity)/2 (%)	Overall accuracy (%)
Ried et al testing set (*n*=23) PMID 17210682	J48	Lymph node status	75.00	81.80	78.40	78.26
Koinuma et al (*n*=17) PMID 16247484	Naïve Bayes	Lymph node status	100.00	75.00	87.50	82.35
Barrier et al (*n*=18) PMID 16091735	Naïve Bayes	Recurrence	100.00	88.90	94.45	94.44
Barrier et al (*n*=50) PMID 16966692	Decision stump	Recurrence	76.00	88.00	82.00	82.00
Barrier et al (*n*=24) PMID 17043639	Naïve Bayes	Recurrence	50.00	100.00	75.00	79.16
NCI 60 U133A (*n*=34)	Threshold selector	Drug response (5-FU)	94.10	76.50	85.30	85.29
NCI 60 U133B (*n*=34)	Threshold selector	Drug response (5-FU)	82.40	76.50	79.45	79.41

3.5 Correlation analysis on CGH and RNA data

3.5.1 Description of the data sets

CGH (Comparative genomic hybridization) data: The array CGH data was available only for 29 of the 73 patient samples. The data consisted of probe name, chromosome name, start and stop coordinates, feature number, and description of the genes.

RNA (Ribonucleic acid) data: The RNA data consisted of 22,464 genes and 73 tumor samples.

3.5.2 Correlation coefficient analysis on the 9-gene signature

This study focused on identifying genes in the 9-gene signature whose cDNA copy number was correlated with the RNA expression data. The CGH data was checked for matching genes with the 9-gene signature and 89 matched probes were found. The same 29 patient samples available in the array CGH data were selected from RNA data. The 9-gene signature and 29 sample RNA expression data versus 9-gene signature and 29 sample CGH data was used to compute the correlation coefficient for each of the genes. After computing the correlation coefficient for each of the matching gene pairs, the genes with absolute value of correlation coefficient > 0.36 were considered to be significant ($p < 0.05$). The gene *ITGB1* satisfied this condition and Table 3.29 shows the details. These genes might be helpful in identifying the regulators of gene expression.

Table 3.29 Genes with correlation coefficient > 0.36 in the CGH versus RNA data.

Probe name	Chromosome name	Start location	Stop location	Feature number	Gene symbol	Correlation coefficient
A_14_P128618	10	33274603	33274662	22571	ITGB1	0.4767
A_14_P201824	10	33284494	33284553	22255	ITGB1	0.4321

3.5.3 Genome wide correlation analysis on CGH and RNA data

Our aim was to identify the cDNA copy numbers of the genes that were correlated with RNA expression data. The 29 tumor samples that were available in the CGH data were selected from the RNA data for the analysis. The genes in the CGH data were matched with the genes in the RNA data. Correlation coefficient was calculated for each of these matched gene pairs across the 29 tumor samples. The obtained correlation coefficients for each of the genes were converted to their absolute values, and all genes which with correlation coefficient values < 0.36 were removed. From the remaining set of genes, 3 or more different consecutive genes were selected. Table 3.30 shows in detail the identified genes. The chromosome locations of these genes might be important in identifying the regulators of gene expression.

Table 3.30 Details of the genes identified by genome-wide correlation analysis.

Chromosome name	Start	Stop	Feature number	Number of consecutive genes	Consecutive genes (count)
1	11788695	11788743	144635	6	MTHFR (2) CLCN6 (4)
1	23864785	23864839	162971	5	LYPLA2 (1) GALE (1) MGCL (3)
1	35990614	35990665	51144	6	EIF2C4 (1) EIF2C1 (5)
1	94617936	94617995	64611	10	ABCD3 (8) F3 (2)
1	1.13E+08	1.13E+08	79186	3	MOV10 (1) RHOC (2)
9	1.09E+08	1.09E+08	4486	5	ACTL7A (1) IKBKAP (4)
9	1.24E+08	1.24E+08	1752	6	NEK6 (1) PSMB7 (5)
9	1.37E+08	1.37E+08	81538	3	MGC14141 (2) KIAA1984 (1)
11	18374755	18374806	146020	3	LDHA (2) LDHC (1)
12	54789993	54790052	78424	3	PA2G4 (1) RPL41 (2)
14	75183472	75183531	75020	3	C14orf58 (1) C14orf1 (2)
14	95918447	95918506	126315	6	C14orf129 (1) AK7 (5)
16	23557994	23558053	42174	3	FLJ21816 (1) MGC3248 (2)
17	7232718	7232766	159975	3	TNK1 (2) PLSCR3 (1)
17	18160066	18160125	44369	5	SMCR8 (1) SHMT1 (4)
17	35087534	35087593	67698	8	PERLD1 (2) ERBB2 (5) C17orf37 (1)
20	17499809	17499868	161401	5	DSTN (4) RRBP1 (1)
20	32972022	32972071	118483	5	ACAS2 (2) GSS (3)
20	34697094	34697153	63618	8	SLA2 (2) NDRG3 (6)
20	43442415	43442474	29649	7	C20orf35 (5) PIGT (2)
20	60315571	60315630	19595	4	ADRM1 (2) LAMA5 (2)
21	26011625	26011684	70488	6	ATP5J (3) GABPA (3)
X	48506958	48507011	181787	3	TIMM17B (2) PQBP1 (1)

3.6 Summary

In this chapter, we identified a 9-gene signature to predict lymph node metastasis in colon cancer patients based on the microarray gene expression data. This was achieved by, firstly preprocessing the data to discard undifferentiated genes using the t-test, secondly replacing missing values with the *k*-nearest neighbors algorithm, and thirdly applying variable selection using random forests. In the next step, the genes without differential expression in lymph node negative versus positive tumors were removed in order to retain only the discriminator genes and obtained the 9-gene signature. The Kaplan-Meier plots of the 9-gene signature on Ried et al data ($n=73$) generated significant patient stratification into, low-risk and high-risk groups (log-rank tests, $p < 0.05$), with distinct relapse-free survival (RFS) and overall survival (OS). Out of the 73 patients in Ried et al data, 26 patients remained relapse-free for more than 5 years and 10 patients experienced relapse within 5 years after surgery. In these patients, the Cox model had a sensitivity of 60.0% and a specificity of 92.3%. The 9-gene lymph node signature was cross validated on independent colon cancer data sets. The drug response to 5-FU (fluorouracil) on the NCI-60 cell line data was predicted. Our results showed that it is feasible to predict the lymph node status of the patients with the 9-gene signature and it might be used for tailored treatments for patients in the high-risk group. Correlation analysis was done between the CGH and RNA data using the 14 gene signature and the gene ITGB1 was identified which exhibited strong relationship between the two groups. Genome wide correlation analysis was done to identify DNA copy numbers of the genes that were correlated with RNA expression data. These results might be useful in identifying the regulators of gene expression.

Chapter 4

Prediction models for recurrence in colon cancer

4.1 Introduction

Recurrence or relapse is the reappearance of a tumor or the return of symptoms after treating for cancer. Postoperative treatment given to Duke's stage B and Duke's stage C colon cancer patients is highly debatable (24). It is uncertain whether adjuvant chemotherapy should be given to Duke's stage B patients because not all the patients benefit from it. So, there is a need to identify patients at high-risk of recurrence who would develop relapse in the Duke's B group so that they can be given aggressive treatment, and patients at low-risk of recurrence would be spared from the invasive treatment. Our study aims at identifying patients at low and high-risks of recurrence by building prognostic models for stage II (Duke's stage B) and stage III (Duke's stage C) colon cancer patients.

The training set comprised of 36 patients (10 patients having recurrence within 5 years after surgery and 26 patients remaining relapse free for more than 5 years). The remaining 37 patients formed the testing set. The missing values in the gene expression data were replaced using the *k*-nearest neighbors algorithm with *k*=30. A combinatorial scheme was used to identify biomarkers predicting the recurrence. In the first step, variable selection using random forests was applied on the training set and a 4-gene subset was obtained. In the second step, InfoGain feature selection technique was applied to further reduce the dimensionality by dropping lower ranked genes, and hence obtained the 3-gene signature. The same procedure was repeated again by replacing missing values in the preprocessed data with *k*-nearest neighbors algorithm (*k*=20) and obtained the 5-gene signature.

The performance of these signatures was evaluated by cross validation on independent colon cancer data sets. The discriminatory powers of the identified gene signatures were evaluated by the time-dependent ROC technique, and these signatures could effectively stratify patients into low-risk and high-risk groups. Prediction models were built with the 3-gene signature and the 5-gene signature using classifiers in Weka software to predict recurrence in patients from the testing set. These gene signatures were also cross validated on independent colon cancer datasets to evaluate their performance.

The remainder of this chapter is organized as follows. Section 4.2 introduces our study design and describes the experiment in detail. Section 4.3 describes the validation results. Section 4.4 discusses the study design of 5-gene signature and describes the experiment in detail. Section 4.5 describes the validation results. Section 4.6 compares the 3-gene and 5-gene signatures, and Section 4.7 provides a summary of this chapter.

4.2 Study Design for the 3-gene signature

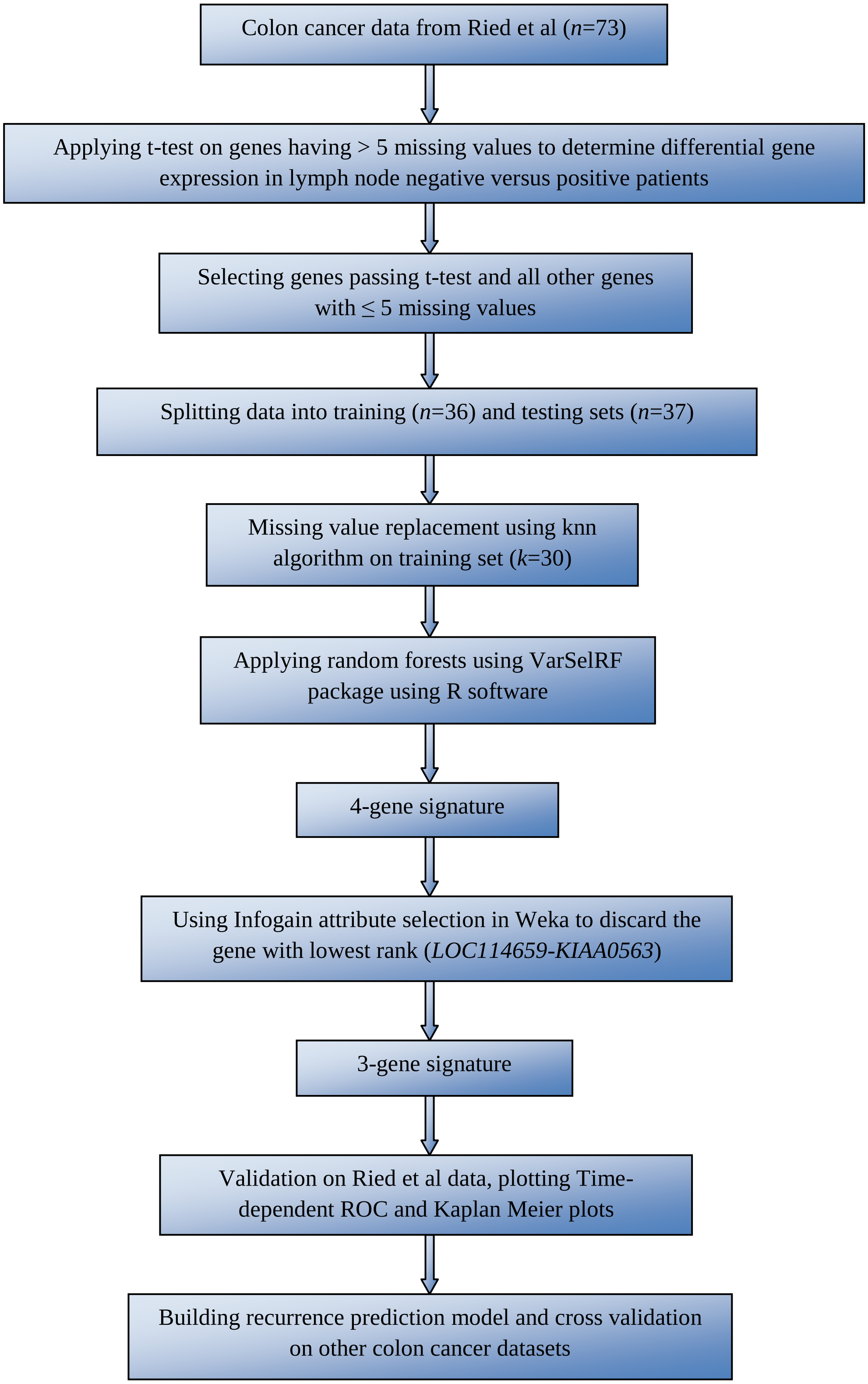
Colon cancer data from Ried et al (*n*=73)
Applying t-test on genes having > 5 missing values to determine differential gene expression in lymph node negative versus positive patients
Selecting genes passing t-test and all other genes with ≤ 5 missing values
Splitting data into training (*n*=36) and testing sets (*n*=37)
Missing value replacement using knn algorithm on training set (*k*=30)
Applying random forests using VarSelRF package using R software
4-gene signature
Using Infogain attribute selection in Weka to discard the gene with lowest rank (*LOC114659-KIAA0563*)
3-gene signature
Validation on Ried et al data, plotting Time-dependent ROC and Kaplan Meier plots
Building recurrence prediction model and cross validation on other colon cancer datasets

Figure 4.1 Block diagram of the study for 3-gene signature.

4.2.1 Experimental procedure

Data Source The colon cancer microarray data from Ried et al. (13) contained 22,464 genes and 73 patient samples, all of them treated for primary adenocarcinomas of the colon. Of these 33 tumor samples were stage II (lymph node negative) and 40 tumor samples were stage III (lymph node positive).

Log Ratio Every spot on the microarray provides two intensity values each of them associated with a specific channel. Dividing one intensity by the other gives the expression ratio. We used log ratios as they are lot easier to work with than the regular ratios. The log ratio (532/635) was considered for this analysis. It is the log (base 2) transformation of the ratio of medians at wavelengths of 532nm and 635 nm.

Data Preprocessing-t test We investigated whether the observed difference between the two groups (node positive versus negative) represents a real difference in the total study population from which the sample was drawn, or whether it just occurred by chance (due to sampling variation), by using t-test. The number of missing values for each gene was found and t-test was done on genes with more than 5 missing values to evaluate the difference in the proportions of missing values in node positive versus negative groups. The genes passing the t-test ($p < 0.05$, two-sided) along with all other genes were selected for further processing. A total of 10,220 genes satisfied this condition.

Training dataset The training dataset consisted of the patient samples having recurrence within 5 years after surgery, survival time ≥ 60 months selected based on the clinical data available for the dataset. The expression data of the 10,220 genes and 36 patients, constituted the training set. The remaining patients formed the testing set.

Missing value replacement The training dataset contained missing values. They were replaced using the EMV [28] package in R software with k=30. This technique estimates the missing values based on the k-nearest neighbors algorithm. This algorithm selects the k nearest rows that do not contain any missing

[28] http://cran.r-project.org/web/packages/EMV/index.html

values to the one containing at least one missing value based on the Euclidian distance. Then the missing values are replaced by the average of the neighbors.

Biomarker identification VarSelRF[29] package in R was used in a series of steps on the training dataset to find the important features. The recurrence status was used as the class variable. In the first step, a forest with N trees was built and the features were ranked according to the importance of the variables. In the second step, 20% of the variables that were least important were removed and a new forest was constructed with K trees. This step was repeated till there were two genes left. In the experiment, a value of N = 2000 and K =1000 were considered, because a large number of trees in the initial forests is likely to produce stable importance measures (23). After fitting all forests, the OOB error rates from all the fitted random forests were examined and a set of 4 genes leading to the smallest error rate were selected. InfoGain attribute selection technique was applied to drop the least ranked gene, *LOC114659--KIAA0563*, giving us the 3-gene signature. Table 4.1 shows the 3-gene signature.

Table 4.1 The 3-gene signature for predicting colon cancer recurrence.

GENE NAME	ID
LRRC14-leucine rich repeat containing	H200014103
E2F2-E2F transcription factor 2 (E2F2)	H200012309
SLC25A5-solute carrier family 25	H200006643

[29] http://cran.r-project.org/web/packages/varSelRF/index.html

4.3 Results

4.3.1 Building prediction model using Weka

The training set consisted of expression data of the 3-gene signature in 36 patients (10 patients having recurrence within 5 years after surgery and 26 patients having survival time more than 5 years without recurrence). The remaining patients formed the testing set. Weka software was used for 10 fold cross validation on the training dataset. Different classification schemes in Weka were applied to this dataset to find the best scheme. Table 4.2 shows the top five classifiers including LWL, AD Tree, Multialyer perceptron, IB1, and Logistic regression based on their prediction accuracies. LWL classifier performed better than the other classifiers. It had a sensitivity of 80.00%, a specificity of 96.20%, and an overall accuracy of 91.66%. Table 4.3 shows the confusion matrix for LWL classifier. The difference in overall accuracy between LWL and other classifiers was not statistically significant due to the small sample size. The LWL classifier model was saved and used to predict class (recurrence/no recurrence) for patients in the testing set. Table 4.4 shows the predicted class for patients in the testing set using the LWL model and compares it with the class predictions obtained from the Cox model.

Table 4.2 Comparison of accuracies obtained from different classifiers for predicting recurrence using the 3-gene signature. The improved overall accuracy of the prediction with the LWL classifier compared with other methods was assessed by significance testing (*N* = 36).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity +Specificity)/2 (%)	Overall accuracy (%)	*P*-value
LWL	**80.00**	**96.20**	**88.10**	**91.66**	
AD Tree	60.00	96.20	78.10	83.33	< 0.14
Multilayer perceptron	60.00	92.30	76.15	83.33	< 0.14
IB1	40.00	92.30	66.15	77.77	< 0.06
Logistic regression	30.00	96.20	63.10	77.77	< 0.06

Table 4.3 Confusion matrix obtained from the LWL classifier for predicting recurrence using the 3-gene signature.

Actual/Predicted	**a (recurrence)**	**b (no recurrence)**
a (recurrence)	8	2
b (no recurrence)	1	25

Table 4.4 Predicting recurrence in patients from the testing dataset using the 3-gene signature.

Serial No	Patient Number	Prediction by LWL	Prediction by Cox model	Match
1	CC-P1	No recurrence	No recurrence	Y
2	CC-P2	No recurrence	No recurrence	Y
3	CC-P4	No recurrence	No recurrence	Y
4	CC-P7	Recurrence	Recurrence	Y
5	CC-P8	No recurrence	No recurrence	Y
6	CC-P10	No recurrence	No recurrence	Y
7	CC-P13	No recurrence	No recurrence	Y
8	CC-P18	Recurrence	Recurrence	Y
9	CC-P20	No recurrence	No recurrence	Y
10	CC-P21	No recurrence	No recurrence	Y
11	CC-P22	No recurrence	No recurrence	Y
12	CC-P23	No recurrence	No recurrence	Y
13	CC-P25	No recurrence	No recurrence	Y
14	CC-P28	No recurrence	No recurrence	Y
15	CC-P29	Recurrence	Recurrence	Y
16	CC-P31	Recurrence	No recurrence	-
17	CC-P34	No recurrence	No recurrence	Y
18	CC-P35	No recurrence	Recurrence	-
19	CC-P37	No recurrence	No recurrence	Y
20	CC-P38	No recurrence	Recurrence	-
21	CC-P40	No recurrence	No recurrence	Y
22	CC-P42	No recurrence	No recurrence	Y
23	CC-P44	No recurrence	No recurrence	Y
24	CC-P46	No recurrence	Recurrence	-
25	CC-P47	No recurrence	No recurrence	Y
26	CC-P48	No recurrence	No recurrence	Y
27	CC-P50	Recurrence	Recurrence	Y
28	CC-P51	Recurrence	Recurrence	Y
29	CC-P55	No recurrence	Recurrence	-
30	CC-P56	No recurrence	Recurrence	-
31	CC-P60	Recurrence	No recurrence	-
32	CC-P62	No recurrence	No recurrence	Y
33	CC-P66	No recurrence	No recurrence	Y
34	CC-P68	Recurrence	Recurrence	Y
35	CC-P70	Recurrence	No recurrence	-
36	CC-P71	Recurrence	No recurrence	-
37	CC-P72	Recurrence	No recurrence	-

4.3.2 Plotting Kaplan-Meier curves based on the patient subgroups obtained from LWL prediction model on data from Ried et al (*n=73*) using the 3-gene signature

The LWL recurrence prediction model discussed in the previous section generated two subgroups of patients, no recurrence and recurrence, on the training and testing data sets. Kaplan-Meier curves were plotted based on the expression data of 3-gene signature in the 73 patient samples (Ried et al data) and the patient subgroups obtained from LWL prediction model. The Kaplan-Meier plots generated significant patient stratification into no recurrence and recurrence groups ($p < 0.05$, $n=73$, log-rank tests), with distinct relapse-free survival. Figure 4.2 shows the survival probabilities for each of the patient subgroups for relapse-free survival.

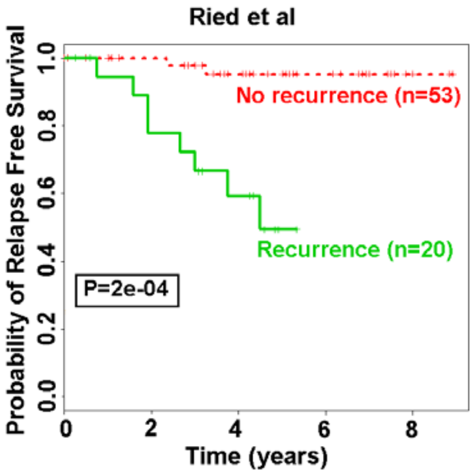

Figure 4.2 Kaplan-Meier plots on data from Ried et al (*n=73*) for relapse-free survival using the 3-gene signature based on the patient subgroups obtained from LWL recurrence prediction model.

4.3.3 Time-dependent ROC analyses data from Ried et al (*n=73*)

To explore whether the 3-gene recurrence signature could predict patient disease-free survival and overall survival, the survival and status information along with the expression data of the 3 genes were used for getting the time-dependent ROC curves. The accuracy of 5-year relapse-free survival prediction using these 3 genes is 0.80 and 5-year overall survival prediction is 0.79, as represented by the AUC.

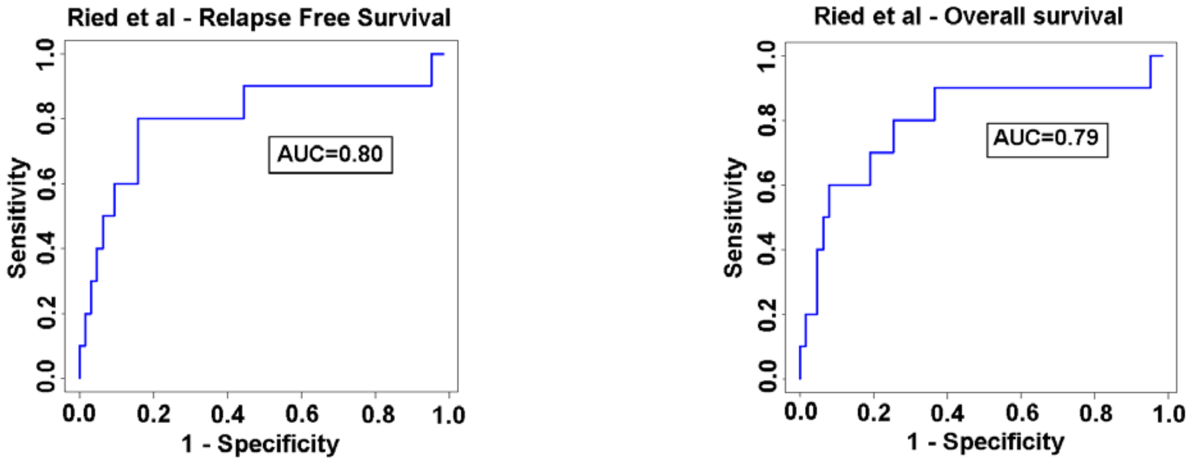

Figure 4.3 Time-dependent ROC plots on data from Ried et al (*n*=*73*) for relapse-free survival and overall survival using the 3-gene signature.

4.3.4 Kaplan-Meier analyses on data from Ried et al (*n*=*73*)

The Cox model based on the expression of the 3-gene signature was used to get recurrence risk scores for all the 73 patient samples. The choices for choosing a cut-off value for patient stratification are the peak value from histogram, mean risk score or median risk score. In this analysis, the peak value from histogram was chosen as cut-off as it resulted in best patient stratification. Cut-off values of 2.0 and 1.0 were chosen for relapse-free survival and overall survival, respectively. The *pamr* package in R was used to plot the relapse-free survival probability of low-risk and high-risk groups. The low-risk and high-risk groups had distinct relapse-free survival (*p* = *1e-04*, *n*=*73*, log-rank tests). The low-risk and high-risk groups had distinct overall survival (*p* = *0.011*, *n*=*73*, log-rank tests).

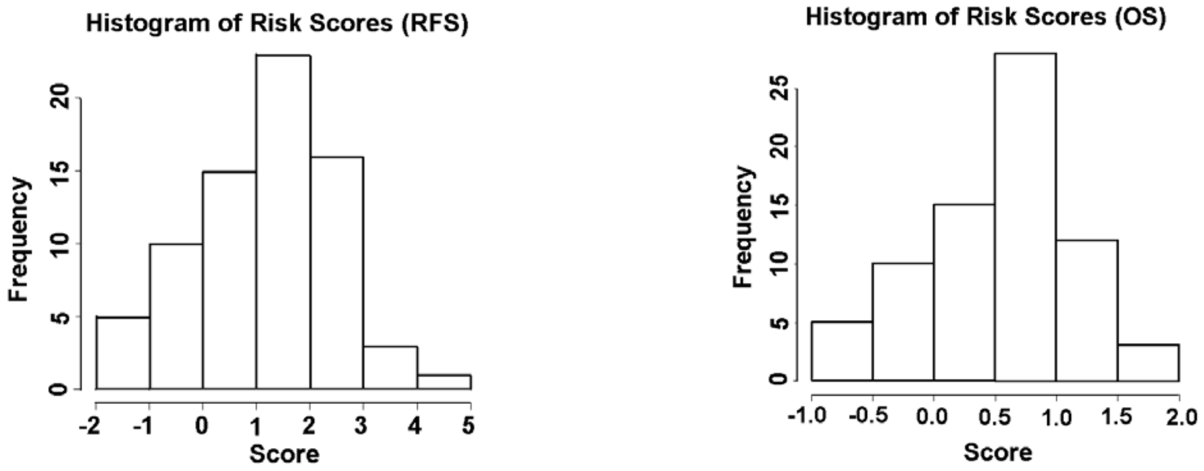

Figure 4.4 Histograms of risk scores obtained from Cox model for relapse-free survival and overall survival using the 3-gene signature

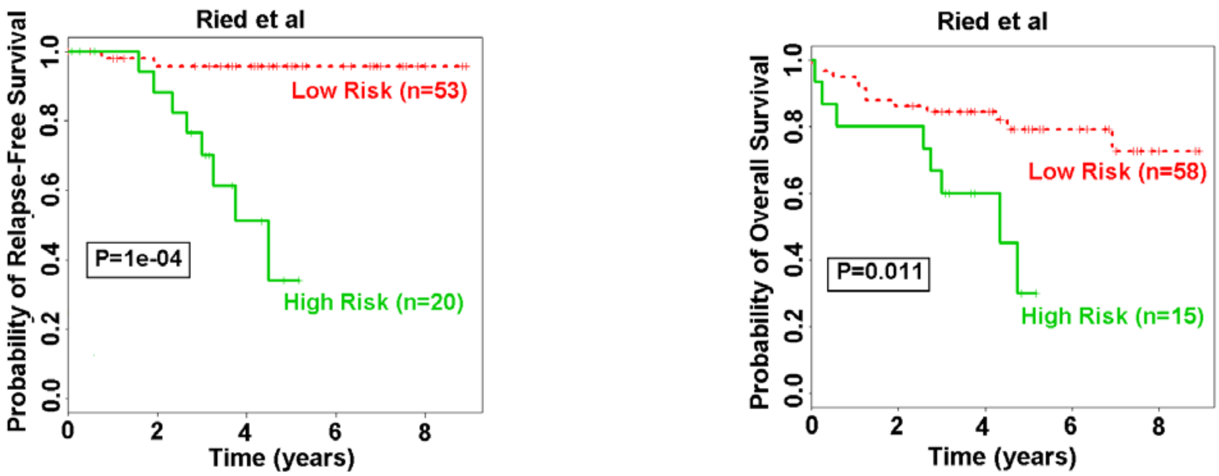

Figure 4.5 Kaplan-Meier plots on data from Ried et al (*n=73*) for relapse-free survival and overall survival using the 3-gene signature

Out of the 73 patients in the colon cancer data from Ried et al, 26 patients remained relapse-free for more than 5 years and 10 patients had recurrence within 5 years after surgery. To test the performance of the 3-gene signature, the subgroups obtained for the above group of 36 patients from the Cox model were compared with their actual clinical outcomes. Table 4.5 shows the different parameters obtained from Cox model using the 3-gene signature, for relapse-free survival and overall survival, respectively. Tables

4.6 and 4.7, show the comparison of predicted clinical outcome for patients with their actual follow-up information, for relapse-free survival and overall survival, respectively. The Cox model had a sensitivity of 80.0%, a specificity of 96.1%, and an overall accuracy of 91.7%, for predicting relapse-free survival. In predicting overall survival, it had a sensitivity of 40.0%, a specificity of 95.9%, and an overall accuracy of 70.4%.

Table 4.5 Different parameters obtained from Cox model using the 3-gene signature for relapse-free survival and overall survival

	Relapse-free survival					Overall survival				
Gene Symbol	**coef**	**exp (coef)**	**se (coef)**	**z-score**	***p*-value**	**coef**	**exp (coef)**	**se (coef)**	**z-score**	***p*-value**
LRRC14	0.449	1.647	0.529	0.943	0.350	0.213	1.238	0.376	0.568	0.570
E2F2	-1.500	0.223	0.586	-2.561	0.010	-0.666	0.513	0.335	-1.988	0.047
SLC25A5	0.375	1.455	0.216	1.738	0.082	0.081	1.085	0.179	0.455	0.650

Table 4.6 Comparison of the sub groups predicted from the Cox model using the 3-gene signature with the actual subgroups for relapse-free survival

	Recurrence	No recurrence	Sensitivity (%)	Specificity (%)	Overall accuracy (%)
Recurrence	8	2	80.0	96.1	91.7
No recurrence	1	25			

Table 4.7 Comparison of the sub groups predicted from the Cox model using the 3-gene signature with the actual subgroups for overall survival

	Death	Alive	Sensitivity (%)	Specificity (%)	Overall accuracy (%)
Death	8	12	40.0	95.9	70.4
Alive	1	23			

The Cox model was used for stratifying all the 73 patient samples in Ried et al data into low-risk and high-risk groups, based on the 3-gene signature. Out of the 73 patients, a total of 37 patients had no recurrence with survival times less than 5 years. Twenty-nine patients had overall survival times less than 5 years without any event (death). The relapse outcome for the 37 patients and the overall survival outcome for the 29 patients is currently unknown. Table 4.8 shows the prospective prognostic predictions of these patients obtained from the Cox model for relapse-free survival and overall survival, respectively.

The follow-up information for these patients is being collected. When it becomes available in the future, the predictions can be validated with it.

Table 4.8 Patient subgroups obtained from the Cox model for relapse-free survival using the 3-gene signature

Serial Number	Patient ID	Predicted group by Cox model (RFS)	Patient ID	Predicted group by Cox model (OS)
1	CC-P1	Low Risk	CC-P1	Low Risk
2	CC-P2	Low Risk	CC-P4	Low Risk
3	CC-P4	Low Risk	CC-P7	Low Risk
4	CC-P7	High Risk	CC-P8	Low Risk
5	CC-P8	Low Risk	CC-P9	Low Risk
6	CC-P10	Low Risk	CC-P11	Low Risk
7	CC-P13	Low Risk	CC-P13	Low Risk
8	CC-P18	High Risk	CC-P16	Low Risk
9	CC-P20	Low Risk	CC-P18	High Risk
10	CC-P21	Low Risk	CC-P20	Low Risk
11	CC-P22	Low Risk	CC-P21	Low Risk
12	CC-P23	Low Risk	CC-P22	Low Risk
13	CC-P25	Low Risk	CC-P25	Low Risk
14	CC-P28	Low Risk	CC-P28	Low Risk
15	CC-P29	High Risk	CC-P31	Low Risk
16	CC-P31	Low Risk	CC-P35	High Risk
17	CC-P34	Low Risk	CC-P36	High Risk
18	CC-P35	High Risk	CC-P37	Low Risk
19	CC-P37	Low Risk	CC-P38	Low Risk
20	CC-P38	High Risk	CC-P40	Low Risk
21	CC-P40	Low Risk	CC-P48	Low Risk
22	CC-P42	Low Risk	CC-P50	High Risk
23	CC-P44	Low Risk	CC-P51	High Risk
24	CC-P46	High Risk	CC-P60	Low Risk
25	CC-P47	Low Risk	CC-P62	Low Risk
26	CC-P48	Low Risk	CC-P66	Low Risk
27	CC-P50	High Risk	CC-P71	Low Risk
28	CC-P51	High Risk	CC-P72	Low Risk
29	CC-P55	High Risk	CC-P73	High Risk
30	CC-P56	High Risk		
31	CC-P60	Low Risk		
32	CC-P62	Low Risk		
33	CC-P66	Low Risk		
34	CC-P68	High Risk		
35	CC-P70	Low Risk		
36	CC-P71	Low Risk		
37	CC-P72	Low Risk		

4.3.5 Independence of 3-gene recurrence signature of tumor stage

Improved prediction of recurrence can profoundly affect clinical decisions. However, following the current clinical guidelines, few of the lymph node-negative patients (stage II) are offered adjuvant chemotherapy. Because 25% to 40% of the patients would develop tumor relapse, the prognosis signature can be a powerful tool to select the patients who are at high-risk and ensure that they receive adjuvant treatment. This part of the study was focused on verifying if the recurrence predictions obtained on Ried et al data were statistically significant when validated separately in Stage II and Stage III patients. It was seen that the 3-gene signature could stratify the patients into low-risk and high-risk groups in Stage II and Stage III samples individually with distinct relapse-free survival. The patient subgroups were obtained based on the Cox model. The patients belonging to the low-risk group had higher survival probabilities than those belonging to the high-risk group. Based on the predictions from LWL model using the 3-gene signature, Kaplan-Meier plots were plotted in Stage II and Stage III samples separately. But the patient stratification was not statistically significant and the results were not reported. So it can be said that Cox model is the best model for predicting recurrence using the 3-gene signature. These results confirm that the 3-gene recurrence signature might be applicable to prognostic categorization for the clinical management of colon cancer.

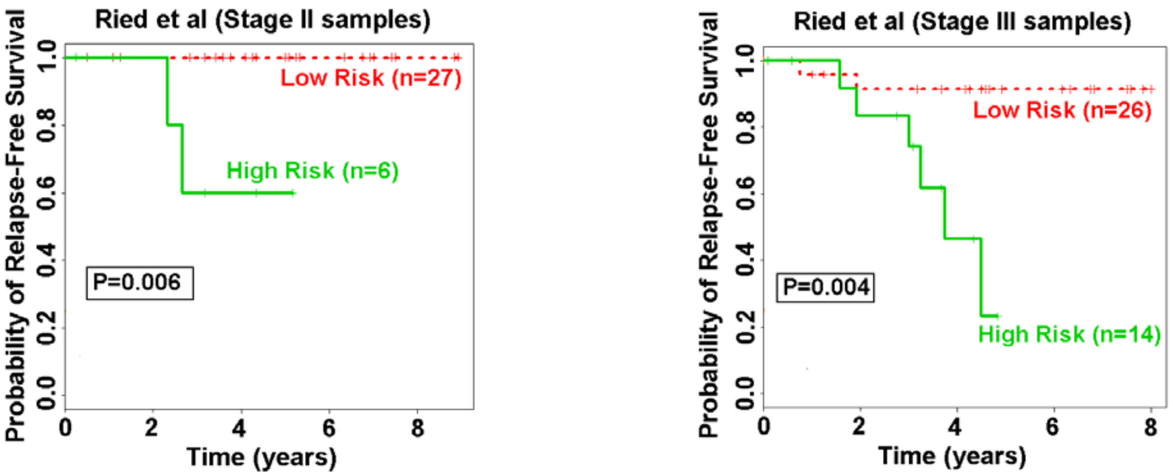

Figure 4.6 The 3-gene signature stratifies patients in Stage II tumors and Stage III tumors into distinct low-risk and high-risk groups for relapse-free survival based on the Cox model.

4.3.6 External validation of the 3-gene signature on other colon cancer data

This part of the study sought to explore the extent to which the 3-gene signature could be used for prediction of lymph node status, recurrence, and drug response in publicly available independent datasets. More than 50 classifiers available in Weka software were tested using a leave-one-out cross validation technique on each of the independent datasets to find a suitable classification scheme for validation. Due to the different number of attributes (matching genes), sample sizes and prediction variables one specific scheme could not be used for validation on all the datasets. Different classifiers had to be employed on the validation datasets to get fair prediction accuracy. As far as possible the same set of classifiers were presented in the comparison tables of validation datasets to provide a fair evaluation of the performance. The exact same set of classifiers could not be compared over all the validation datasets due to poor performances of classifiers on some datasets and good performances on other datasets. The following sections discuss the validation results and comparisons of various classifiers on the independent datasets in detail.

4.3.6.1 Predicting lymph node status by leave-one-out cross validation on data from Koinuma et al. (*n=17*) (PMID 16247484)

The data from Koinuma et al (Affymetrix HG U133 A platform) consisted of 20 patient samples of which 3 patients were Duke's stage D. The Duke's stage D patients were not considered for validation. The search for matching genes with the 3-gene signature was done using the Affymetrix ids. There were 3 matching genes (Table 4.9). The data used for validation consisted of the expression of these 3 genes in the 17 patient samples. Weka software was used for validation and lymph node status (positive/negative) was predicted. Different classification schemes including Multilayer perceptron, Decision stump, Logistic regression, JRip, and *Adaboost M1* were applied to this dataset to find the best scheme. Table 4.10 shows the comparison between Multilayer perceptron and some of the classifiers used for validation on other datasets. Multilayer perceptron classifier performed better than the other classifiers. It had a sensitivity of 71.40%, a specificity of 90.00%, and an overall accuracy of 82.35%. Table 4.11 shows the confusion matrix for Multilayer perceptron classifier. The difference in overall accuracy between Multilayer perceptron and other classifiers was not statistically significant due to the small sample size.

Table 4.9 Matching genes in Koinuma et al data

GENE NAME	ID
LRRC14-leucine rich repeat containing	H200014103
E2F2-E2F transcription factor 2 (E2F2)	H200012309
SLC25A5-solute carrier family 25	H200006643

Table 4.10 Comparison of accuracies obtained from different classifiers for predicting lymph node status using the 3-gene signature. The improved overall accuracy of the prediction with the Multilayer perceptron classifier compared with other methods was assessed by significance testing (*N* = 17).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Multilayer perceptron	**71.40**	**90.00**	**80.70**	**82.35**	
Decision stump	42.90	100.00	71.45	76.47	< 0.33
Logistic regression	57.10	80.00	68.55	70.58	< 0.21
JRip	14.30	100.00	57.15	64.70	< 0.13
Adaboost M1	28.60	80.00	54.30	58.82	< 0.06

Table 4.11 Confusion matrix obtained from the Multilayer perceptron classifier for predicting lymph node status using the 3-gene signature.

Actual/Predicted	a (positive)	b (negative)
a (positive)	5	2
b (negative)	1	9

4.3.6.2 Predicting recurrence by leave-one-out cross validation on data from Barrier et al. (*n=18*) (PMID 16091735)

The data from Barrier et al (PMID 16091735) consisted of 22,283 genes and 18 patient samples. The search for matching genes with the 3-gene signature was done using the Affymetrix ids. There were 3 matching genes (Table 4.12). The data used for validation consisted of the expression of these 3 genes in the 18 patient samples. Weka software was used for validation and recurrence (yes/no) was predicted. Different classification schemes including Threshold selector, AdaboostM1, LWL, Multilayer perceptron, and *IB1* were applied to this dataset to find the best scheme. Table 4.13 shows the comparison between Threshold selector and some of the classifiers used for validation on other datasets. Threshold selector classifier performed better than the other classifiers. It had a sensitivity of 88.90%, a specificity of 77.80%, and an overall accuracy of 83.33%. Table 4.14 shows the confusion matrix for Threshold selector classifier. The difference in overall accuracy between Threshold selector and other classifiers was not statistically significant due to the small sample size.

Table 4.12 Matching genes in Barrier et al data.

GENE NAME	ID
LRRC14-leucine rich repeat containing	H200014103
E2F2-E2F transcription factor 2 (E2F2)	H200012309
SLC25A5-solute carrier family 25	H200006643

Table 4.13 Comparison of accuracies obtained from different classifiers for predicting recurrence using the 3-gene signature. The improved overall accuracy of the prediction with the Threshold selector classifier compared with other methods was assessed by significance testing (*N* = 18).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Threshold selector	**88.90**	**77.80**	**83.35**	**83.33**	
AdaboostM1	77.80	77.80	77.80	77.77	< 0.34
LWL	77.80	77.80	77.80	77.77	< 0.34
Multilayer perceptron	77.80	66.70	72.25	72.22	< 0.22
IB1	66.70	66.70	66.70	66.66	< 0.13

Table 4.14 Confusion matrix obtained from the Threshold selector classifier for predicting recurrence using the 3-gene signature

Actual/Predicted	a (recurrence)	b (no recurrence)
a (recurrence)	8	1
b (no recurrence)	2	7

4.3.6.3 Predicting recurrence by leave-one-out cross validation on data from Barrier et al. (*n=50*) (PMID 16966692)

The data from Barrier et al (PMID 16966692) consisted of 22,283 genes and 50 patient samples. The search for matching genes with the 3-gene signature was done using the Affymetrix ids. There were 3 matching genes (Table 4.15). The data used for validation consisted of the expression of these 3 genes in the 50 patient samples. Weka software was used for validation and recurrence (yes/no) was predicted. Different classification schemes including *IB1*, LWL, Multilayer perceptron, Random committee, and *AdaboostM1* were applied to this dataset to find the best scheme. Table 4.16 shows the comparison between *IB1* and some of the classifiers used for validation on other datasets. *IB1* classifier performed better than the other classifiers. It had a sensitivity of 76.00%, a specificity of 80.00%, and an overall accuracy of 78.00%. Table 4.17 shows the confusion matrix for *IB1* classifier. The difference in overall accuracy between *IB1* and other classifiers was not statistically significant due to the small sample size.

Table 4.15 Matching genes in Barrier et al data.

GENE NAME	ID
LRRC14-leucine rich repeat containing	H200014103
E2F2-E2F transcription factor 2 (E2F2)	H200012309
SLC25A5-solute carrier family 25	H200006643

Table 4.16 Comparison of accuracies obtained from different classifiers for predicting recurrence using the 3-gene signature. The improved overall accuracy of the prediction with the *IB1* classifier compared with other methods was assessed by significance testing (*N* = 50).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
IB1	**76.00**	**80.00**	**78.00**	**78.00**	
LWL	68.00	84.00	76.00	76.00	< 0.41
Multilayer perceptron	80.00	68.00	74.00	74.00	< 0.32
Random committee	68.00	72.00	70.00	70.00	< 0.18
AdaboostM1	56.00	80.00	68.00	68.00	< 0.12

Table 4.17 Confusion matrix obtained from the *IB1* classifier for predicting recurrence using the 3-gene signature.

Actual/Predicted	**a (no recurrence)**	**b (recurrence)**
a (no recurrence)	20	5
b (recurrence)	6	19

4.3.6.4 Predicting the response of cell lines in NCI-60 (U133A GCRMA) data (*n=34*) by leave-one-out cross validation on data

This dataset[30] consisted of 21,225 genes and 60 cell lines (41). Our focus was on the sensitive and resistant cell lines, so cell lines with intermediate response were not considered for validation. A total of 34 cell lines (17 sensitive and the other 17 resistant to the drug 5-FU) were used in validation. The search for matching genes was done using the gene symbols. There were 3 matching genes (Table 4.18). The data used for validation consisted of the expression of these 3 genes in the 34 cell lines. Weka software was used for validation and the response (sensitive/resistant) for the drug 5-FU (fluorouracil) was predicted. Different classification schemes including Threshold selector, Multilayer perceptron, *IB1*, LWL, and Logistic regression were applied to this dataset to find the best scheme. Table 4.19 shows the comparison between Threshold selector and some of the classifiers used for validation on other datasets. Threshold selector classifier performed better than the other classifiers. It had a sensitivity of 94.10%, a specificity of 88.20%, and an overall accuracy of 91.17%. Table 4.20 shows the confusion matrix for Threshold selector classifier. The difference in overall accuracy between Threshold selector and Multilayer perceptron (p < 0.01), *IB1* (p < 0.01), LWL (p < 0.01), Logistic regression (p < 0.01) was statistically significant.

Table 4.18 Matching genes in NCI-60 U133A data.

GENE NAME	ID
LRRC14-leucine rich repeat containing	H200014103
E2F2-E2F transcription factor 2 (E2F2)	H200012309
SLC25A5-solute carrier family 25	H200006643

[30] http://discover.nci.nih.gov/cellminer/loadDownload.do

Table 4.19 Comparison of accuracies obtained from different classifiers for predicting drug response using the 3-gene signature. The improved overall accuracy of the prediction with the Threshold selector classifier compared with other methods was assessed by significance testing (*N* = 34).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Threshold selector	**94.10**	**88.20**	**91.15**	**91.17**	
Multilayer perceptron	70.60	41.20	55.90	55.88	< 0.01
IB1	64.70	47.10	55.90	55.88	< 0.01
LWL	82.40	29.40	55.90	55.88	< 0.01
Logistic regression	35.30	58.80	47.05	47.05	< 0.01

Table 4.20 Confusion matrix obtained from the Threshold selector classifier for predicting drug response using the 3-gene signature.

Actual/Predicted	**a (sensitive)**	**b (resistant)**
a (sensitive)	16	1
b (resistant)	2	15

4.3.6.5 Predicting the response of cell lines in NCI-60 (U133B GCRMA) data (*n=34*) by leave-one-out cross validation on data

This dataset[31] consisted of 17910 genes and 60 cell lines (41). Our focus was on the sensitive and resistant cell lines, so cell lines with intermediate response were not considered for validation. A total of 34 cell lines (17 sensitive and the other 17 resistant to the drug 5-FU) were used in validation. The search for matching genes was done using the gene symbols. There was 1 matching gene (Table 4.21). The data used for validation consisted of the expression of this gene in the 34 cell lines. Weka software was used for validation and the response (sensitive/resistant) for the drug 5-FU (fluorouracil) was predicted. Different classification schemes including Threshold selector, *IB1*, Multilayer perceptron, LWL, and Bagging were applied to this dataset to find the best scheme. Table 4.22 shows the comparison between Threshold selector and some of the classifiers used for validation on other datasets. Threshold selector classifier performed better than the other classifiers. It had a sensitivity of 88.23%, a specificity of

[31] http://discover.nci.nih.gov/cellminer/loadDownload.do

88.23%, and an overall accuracy of 88.23%. Table 4.23 shows the confusion matrix for Threshold selector classifier. The difference in overall accuracy between Threshold selector and *IB1* (*p* < 0.01), Multilayer perceptron (*p* < 0.01), LWL (*p* < 0.01), Bagging (*p* < 0.01) was statistically significant.

Table 4.21 Matching genes in NCI-60 U133B data.

GENE NAME	ID
E2F2-E2F transcription factor 2 (E2F2)	H200012309

Table 4.22 Comparison of accuracies obtained from different classifiers for predicting drug response using the 3-gene signature. The improved overall accuracy of the prediction with the Threshold selector classifier compared with other methods was assessed by significance testing (*N* = 34).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Threshold selector	**88.23**	**88.23**	**88.23**	**88.23**	
IB1	47.10	52.90	50.00	50.00	< 0.01
Multilayer perceptron	11.80	82.40	47.10	47.05	< 0.01
LWL	29.40	58.80	44.10	44.11	< 0.01
Bagging	35.30	47.10	41.20	41.17	< 0.01

Table 4.23 Confusion matrix obtained from the Threshold selector classifier for predicting drug response using the 3-gene signature.

Actual/Predicted	a (sensitive)	b (resistant)
a (sensitive)	15	2
b (resistant)	2	15

4.3.7 Summary of validation results of 3-gene signature

Table 4.24 shows the details of different validation datasets, predicted variables, classifiers used and different accuracies obtained using the 3-gene signature. For each dataset the classifier with the highest overall accuracy was reported.

Table 4.24 Summary of validation results of 3-gene signature on Ried et al data, independent colon cancer datasets and NCI 60 data.

Dataset	Classifier	Predicted variable	Sensitivity (%)	Specificity (%)	(Sensitivity + Specificity)/2 (%)	Overall accuracy (%)
Ried et al training set (*n*=36) PMID 17210682	LWL	Recurrence	80.00	96.20	88.10	91.66
Ried et al training set (*n*=36) PMID 17210682	Cox model	Recurrence	80.00	96.20	88.10	91.66
Koinuma et al (*n*=17) PMID 16247484	Multilayer perceptron	Lymph node status	71.40	90.00	80.70	82.35
Barrier et al (*n*=18) PMID 16091735	Threshold selector	Recurrence	88.90	77.80	83.35	83.33
Barrier et al (*n*=50) PMID 16966692	IB1	Recurrence	76.00	80.00	78.00	78.00
NCI 60 U133A (*n*=34)	Threshold selector	Drug response (5-FU)	94.10	88.20	91.15	91.17
NCI 60 U133B (*n*=34)	Threshold selector	Drug response (5-FU)	88.23	88.23	88.23	88.23

4.4 Study Design for the 5-gene signature

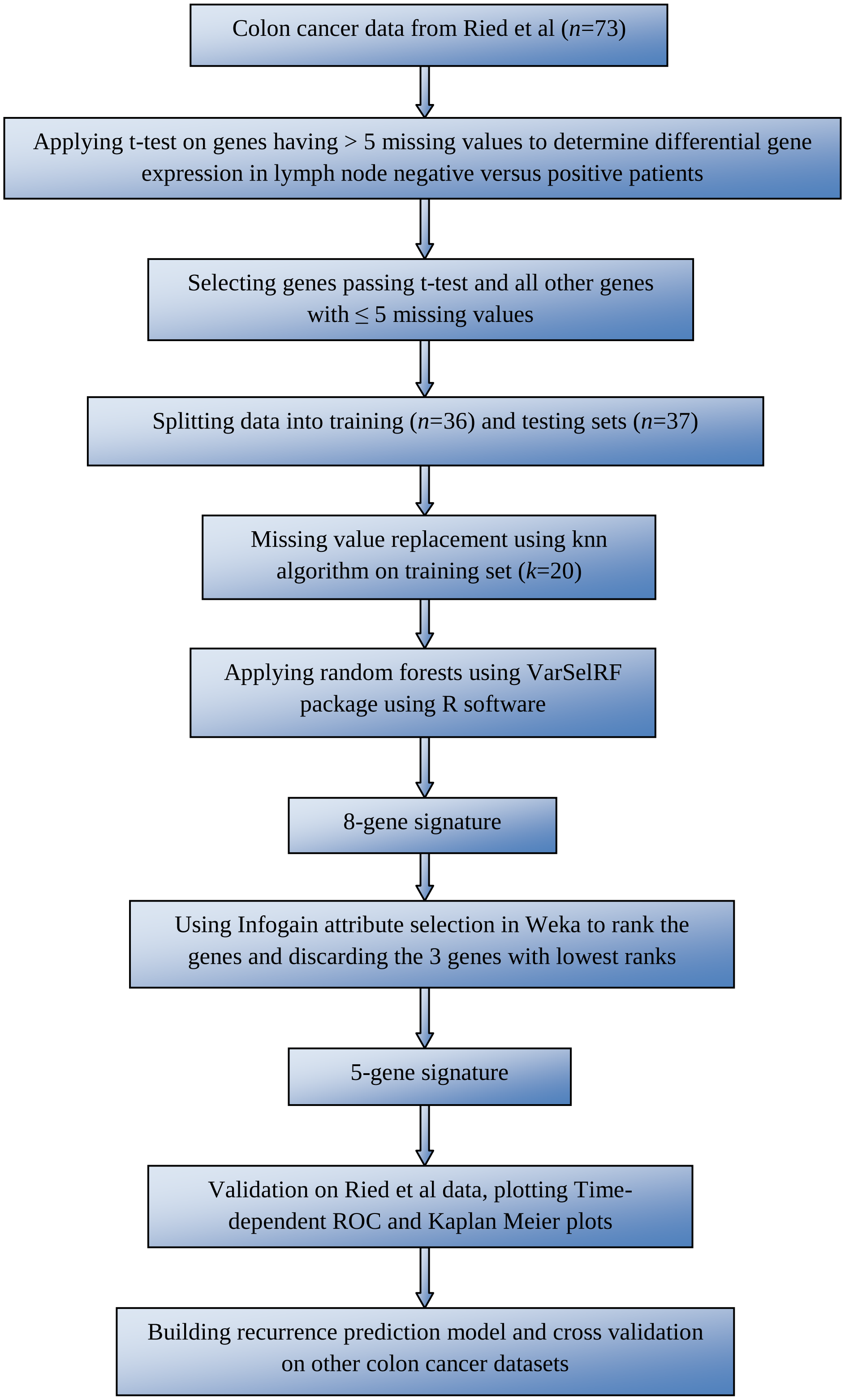
Colon cancer data from Ried et al (*n*=73)
Applying t-test on genes having > 5 missing values to determine differential gene expression in lymph node negative versus positive patients
Selecting genes passing t-test and all other genes with ≤ 5 missing values
Splitting data into training (*n*=36) and testing sets (*n*=37)
Missing value replacement using knn algorithm on training set (*k*=20)
Applying random forests using VarSelRF package using R software
8-gene signature
Using Infogain attribute selection in Weka to rank the genes and discarding the 3 genes with lowest ranks
5-gene signature
Validation on Ried et al data, plotting Time-dependent ROC and Kaplan Meier plots
Building recurrence prediction model and cross validation on other colon cancer datasets

Figure 4.7 Block diagram of the study for 5-gene signature

4.4.1 Experimental procedure

Data Source The colon cancer microarray data from Ried et al. (13) contained 22,464 genes and 73 patient samples, all of them treated for primary adenocarcinomas of the colon. Of these 33 tumor samples were stage II (lymph node negative) and 40 tumor samples were stage III (lymph node positive).

Log Ratio Every spot on the microarray provides two intensity values each of them associated with a specific channel. Dividing one intensity by the other gives the expression ratio. We used log ratios as they are lot easier to work with than regular ratios. The log ratio (532/635) was considered for this analysis which is log (base 2) transformation of the ratio of medians at wavelengths of 532nm and 635 nm.

Data Preprocessing - t test We investigated whether the observed difference between the two groups (node positive versus negative) represents a real difference in the total study population from which the sample was drawn, or whether it just occurred by chance (due to sampling variation), by using t-test. The number of missing values for each gene was found and t-test was done on genes with more than 5 missing values to evaluate the difference in the proportions of missing values in node positive versus negative groups. The genes passing the t-test ($p < 0.05$, two-sided) were included along with all genes having less than 5 missing values for further processing. A total of 10,220 genes satisfied this condition.

Training dataset The training dataset consisted of the patient samples with recurrence, survival time ≥ 60 months selected based on the clinical data available for the dataset. The expression data of the 10,220 genes and 36 patients, constituted the training set. The remaining patients formed the testing set.

Missing value replacement The training dataset contained missing values. They were replaced using the EMV [32] package in R software with k=20. This technique estimates the missing values based on the k nearest neighbors algorithm. This algorithm selects the k nearest rows that do not contain any missing

[32] http://cran.r-project.org/web/packages/EMV/index.html

values to the one containing at least one missing value, based on the Euclidian distance. Then the missing values are replaced by the average of the neighbors.

Biomarker identification VarSelRF[33] package in R was used in a series of steps on the training dataset to find the important features. The recurrence status was used as the class variable. In the first step, a forest with N trees was built and the features were ranked according to the importance of the variables. In the second step, 20% of the variables that were least important were removed and a new forest was constructed with K trees. This step was repeated till there were two genes left. In the experiment, a value of N = 2000 and K =1000 were considered, because a large number of trees in the initial forests is likely to produce stable importance measures (23). After fitting all forests, the OOB error rates from all the fitted random forests were examined and a set of 8 genes leading to the smallest error rate were selected. The InfoGain attribute selection technique was used to drop three least ranked genes *(LOC114659--KIAA0563, cDNA DKFZp564O1172, and NET1)* and obtained the 5-gene signature. Table 4.25 shows the 5-gene signature.

Table 4.25 The 5-gene signature for predicting colon cancer recurrence.

GENE NAME	ID
TPD52L2-tumor protein D52-like2	H200013992
CDNA FLJ44020 fis, clone TESTI4026295	H200020685
ZNF187-zinc finger protein 187 (ZNF187)	H200015602
HSPA14-heat shock 70kDa protein 14	H200018991
SLC25A5-solute carrier family 25	H200006643

[33] http://cran.r-project.org/web/packages/varSelRF/index.html

4.5 Results

4.5.1 Building prediction model using Weka

The training set consisted of expression data of the 5-gene signature in 36 patients (10 patients having recurrence within 5 years after surgery and 26 patients having survival time more than 5 years without recurrence). The remaining patients formed the testing set. 10-fold cross validation was used on the training dataset. Different classification schemes in Weka were applied on the training set to find the best scheme. Table 4.26 shows the top five classifiers including Random Tree, *KStar*, AD Tree, *IB1*, and Multilayer perceptron based on their prediction accuracies. Random Tree classifier performed better than the other classifiers. It had a sensitivity of 70.00%, a specificity of 88.46% and an overall accuracy of 83.33%. Table 4.27 shows the confusion matrix for Random Tree classifier. The difference in overall accuracy between Random Tree and other classifiers was not statistically significant due to the small sample size. The Random Tree classifier model was saved and used to predict class (recurrence/no recurrence) for patients in the testing set. Table 4.28 shows the predicted class for patients in the testing set using the Random Tree prediction model and compares it with the class predictions obtained from the Cox model.

Table 4.26 Comparison of accuracies obtained from different classifiers for predicting recurrence using the 5-gene signature. The improved overall accuracy of the prediction with the Random Tree classifier compared with other methods was assessed by significance testing (*N* = 36).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Random Tree	**70.00**	**88.46**	**79.23**	**83.33**	
KStar	40.00	92.30	66.15	77.77	< 0.28
AD Tree	50.00	84.60	67.30	75.00	< 0.19
IB1	30.00	88.50	59.25	72.22	< 0.13
Multilayer perceptron	30.00	84.60	57.30	69.44	< 0.08

Table 4.27 Confusion matrix obtained from the Random Tree classifier for predicting recurrence using the 5-gene signature.

Actual/Predicted	**a (no recurrence)**	**b (recurrence)**
a (no recurrence)	23	3
b (recurrence)	3	7

Table 4.28 Predicting recurrence in patients from the testing dataset using the 5-gene signature.

Serial. No	Patient Number	Prediction by Random Tree	Prediction by Cox model	Match
1	CC-P1	No recurrence	No recurrence	Y
2	CC-P2	No recurrence	Recurrence	-
3	CC-P4	No recurrence	No recurrence	Y
4	CC-P7	No recurrence	No recurrence	Y
5	CC-P8	No recurrence	Recurrence	-
6	CC-P10	No recurrence	No recurrence	Y
7	CC-P13	No recurrence	No recurrence	Y
8	CC-P18	No recurrence	Recurrence	-
9	CC-P20	No recurrence	Recurrence	-
10	CC-P21	No recurrence	No recurrence	Y
11	CC-P22	Recurrence	Recurrence	Y
12	CC-P23	No recurrence	No recurrence	Y
13	CC-P25	No recurrence	No recurrence	Y
14	CC-P28	No recurrence	Recurrence	-
15	CC-P29	No recurrence	Recurrence	-
16	CC-P31	Recurrence	No recurrence	-
17	CC-P34	No recurrence	Recurrence	-
18	CC-P35	No recurrence	Recurrence	-
19	CC-P37	No recurrence	No recurrence	Y
20	CC-P38	No recurrence	No recurrence	Y
21	CC-P40	No recurrence	No recurrence	Y
22	CC-P42	No recurrence	No recurrence	Y
23	CC-P44	No recurrence	No recurrence	Y
24	CC-P46	No recurrence	No recurrence	Y
25	CC-P47	No recurrence	No recurrence	Y
26	CC-P48	No recurrence	No recurrence	Y
27	CC-P50	No recurrence	No recurrence	Y
28	CC-P51	Recurrence	No recurrence	-
29	CC-P55	No recurrence	No recurrence	Y
30	CC-P56	No recurrence	No recurrence	Y
31	CC-P60	No recurrence	No recurrence	Y
32	CC-P62	No recurrence	No recurrence	Y
33	CC-P66	No recurrence	No recurrence	Y
34	CC-P68	No recurrence	No recurrence	Y
35	CC-P70	No recurrence	Recurrence	-
36	CC-P71	Recurrence	Recurrence	Y
37	CC-P72	No recurrence	Recurrence	-

4.5.2 Plotting Kaplan-Meier curves based on the patient subgroups obtained from Random Tree prediction model on data from Ried et al (*n=73*) using the 5-gene signature

The Random Tree recurrence prediction model discussed in the previous section generated two subgroups of patients, recurrence and no recurrence on the training and testing data sets. Kaplan-Meier curves were plotted based on the expression data of 5-gene signature in the 73 patient samples (Ried et al) and the patient subgroups obtained from Random Tree prediction model. The Kaplan-Meier plots generated significant patient stratification into no recurrence and recurrence groups ($p < 0.05$, $n=73$, log-rank tests), with distinct relapse-free survival. Figure 4.8 shows the survival probabilities for each of the patient subgroups for relapse-free survival.

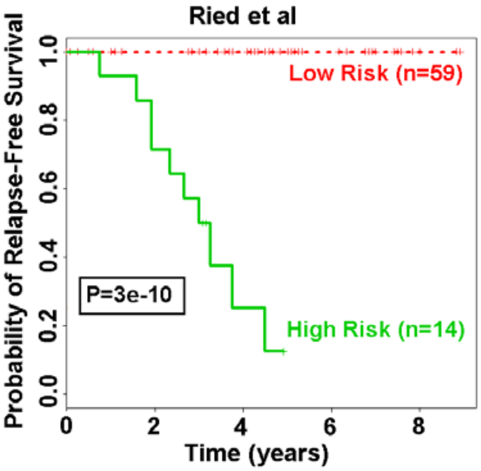

Figure 4.8 Kaplan-Meier plots on data from Ried et al (*n=73*) for relapse-free survival using the 5-gene signature, based on patient subgroups obtained from Random Tree recurrence prediction model.

4.5.3 Time-dependent ROC analyses on data from Ried et al (*n=73*)

To explore whether the 5-gene recurrence signature could predict patient disease-free survival and overall survival, the survival and status information along with the expression data of the 5 genes were used for getting the time-dependent ROC curves. The accuracy of 5-year relapse-free survival prediction using these 5 genes is 0.73 and 5-year overall survival prediction is 0.73, as represented by the AUC.

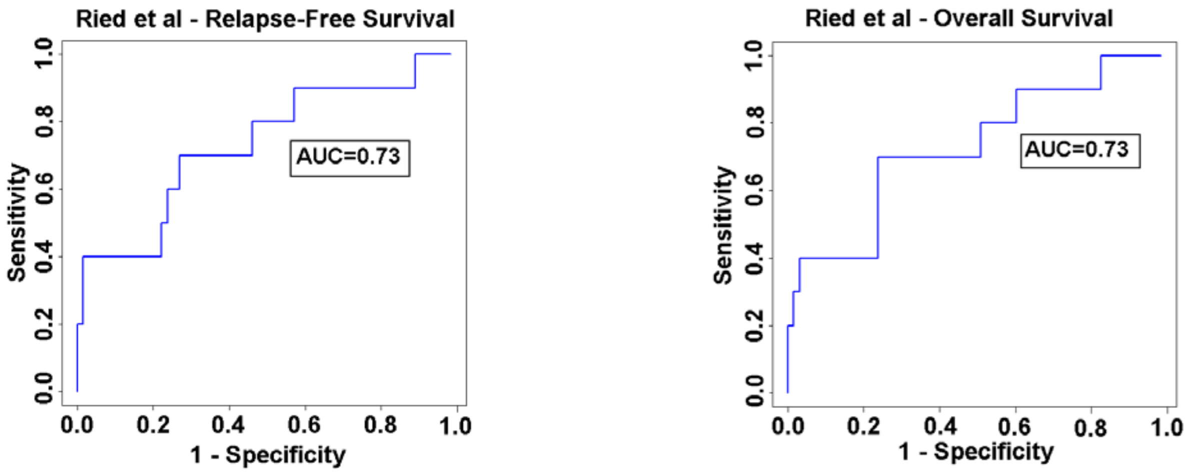

Figure 4.9 Time-dependent ROC plots on data from Ried et al (*n*=*73*) for relapse-free survival and overall survival using the 5-gene signature.

4.5.4 Kaplan-Meier analyses on data from Ried et al (*n*=*73*)

The Cox model based on the expression of the 5-gene signature was used to get recurrence risk scores for all 73 patient samples. The choices for choosing a cut-off value for patient stratification are the peak value from histogram, mean risk score or median risk score. In this analysis, the peak value from histogram was chosen as cut-off as it resulted in best patient stratification. Cut-off values of 0.25 and -0.5 were chosen for relapse-free survival and overall survival, respectively. The *pamr* package in R was used to plot the relapse-free survival probability of low-risk and high-risk groups. The low-risk and high-risk groups had distinct relapse-free survival (*p = 0.01*, *n*=*73*, log-rank tests). The low-risk and high-risk groups had distinct overall survival (*p = 0.04*, *n*=*73*, log-rank tests).

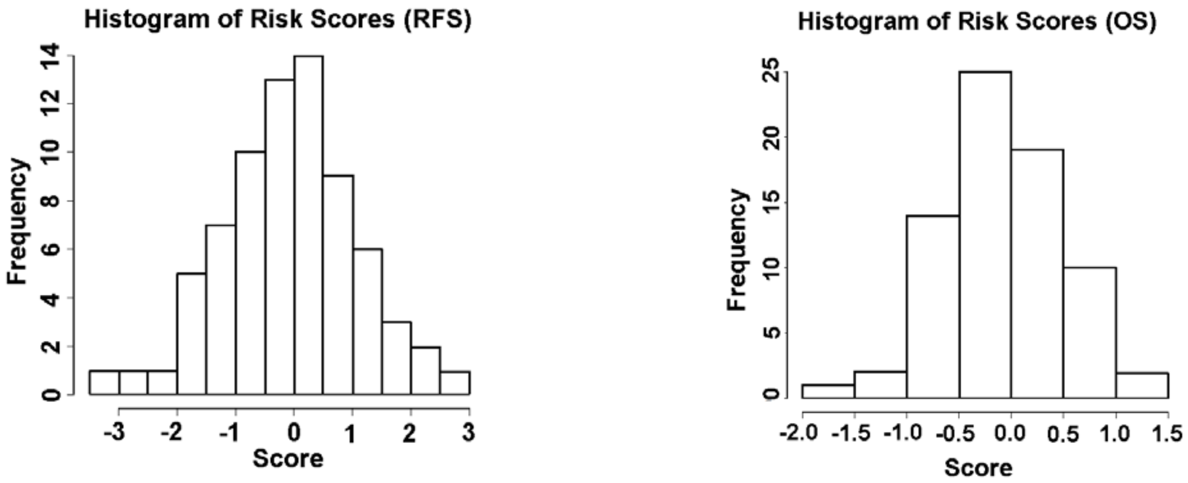

Figure 4.10 Histograms of risk scores obtained from Cox model for relapse-free survival and overall survival using the 5-gene signature.

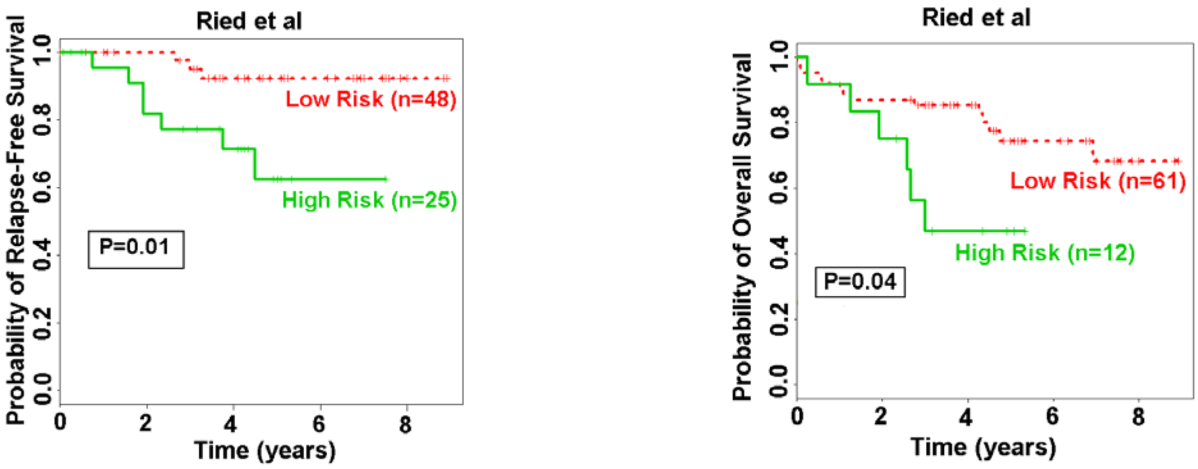

Figure 4.11 Kaplan-Meier plots on data from Ried et al (*n=73*) for relapse-free survival and overall survival using the 5-gene signature.

Out of the 73 patients in the colon cancer data from Ried et al, 26 patients remained relapse-free for more than 5 years and 10 patients had recurrence within 5 years after surgery. To test the performance of the 5-gene signature the subgroups obtained for the above group of 36 patients from the Cox model were compared with their actual clinical outcomes. Table 4.29 shows the different parameters obtained from the Cox model using the 5-gene signature, for relapse-free survival and overall survival, respectively. Tables 4.30 and 4.31, show the comparison of predicted clinical outcome for patients with their actual follow-up information, for relapse-free survival and overall survival, respectively. The Cox model had a sensitivity of 70.0%, a specificity of 80.8%, and an overall accuracy of 77.8%, for predicting relapse-free survival. In predicting overall survival, it had a sensitivity of 30.0%, a specificity of 91.7%, and an overall accuracy of 63.6%.

Table 4.29 Different parameters obtained from Cox model using the 5-gene signature for relapse-free survival and overall survival.

	Relapse-free survival					**Overall survival**				
Gene symbol	**coef**	**exp (coef)**	**se (coef)**	**z-score**	***p*-value**	**coef**	**exp (coef)**	**se (coef)**	**z-score**	***p*-value**
TPD52L2	1.0024	2.725	0.571	1.7548	0.079	0.519	1.680	0.325	1.596	0.11
CDNAFLJ44020	0.6445	1.905	0.483	1.3351	0.180	0.389	1.476	0.278	1.398	0.16
ZNF187	-0.0248	0.975	0.266	-0.0935	0.930	0.081	1.084	0.075	1.079	0.28
HSPA14	0.0749	1.078	0.344	0.2181	0.830	-0.159	0.853	0.214	-0.742	0.46
SLC25A5	-0.1238	0.884	0.349	-0.3547	0.720	-0.092	0.912	0.249	-0.368	0.71

Table 4.30 Comparison of the sub groups predicted from the Cox model using the 5-gene signature with the actual subgroups for relapse-free survival.

	Recurrence	No recurrence	Sensitivity (%)	Specificity (%)	Overall Accuracy (%)
Recurrence	7	3	70.0	80.8	77.8
No recurrence	5	21			

Table 4.31 Comparison of the sub groups predicted from the Cox model using the 5-gene signature with the actual subgroups for overall survival.

	Death	Alive	Sensitivity (%)	Specificity (%)	Overall Accuracy (%)
Death	6	14	30.0	91.7	63.6
Alive	2	22			

The Cox model was used for stratifying all the 73 patient samples in Ried et al data into low-risk and high-risk groups based on the 5-gene signature. Out of the 73 patients, a total of 37 patients had no recurrence with survival times less than 5 years. Twenty-nine patients had overall survival times less than 5 years without any event (death). The relapse outcome for the 37 patients and the overall survival outcome for the 29 patients is currently unknown. Table 4.32 shows the prospective prognostic predictions of these patients obtained from the Cox model for relapse-free survival and overall survival, respectively. The follow-up information for these patients is being collected. When it becomes available in the future, the predictions can be compared with it.

Table 4.32 Patient subgroups obtained from the Cox model for relapse-free survival using the 5-gene signature.

Serial Number	Patient ID	Predicted group by Cox model (RFS)	Patient ID	Predicted group by Cox model (OS)
1	CC-P1	Low Risk	CC-P1	Low Risk
2	CC-P2	High Risk	CC-P4	Low Risk
3	CC-P4	Low Risk	CC-P7	Low Risk
4	CC-P7	Low Risk	CC-P8	Low Risk
5	CC-P8	High Risk	CC-P9	Low Risk
6	CC-P10	Low Risk	CC-P11	High Risk
7	CC-P13	Low Risk	CC-P13	Low Risk
8	CC-P18	High Risk	CC-P16	Low Risk
9	CC-P20	High Risk	CC-P18	Low Risk
10	CC-P21	Low Risk	CC-P20	Low Risk
11	CC-P22	High Risk	CC-P21	Low Risk
12	CC-P23	Low Risk	CC-P22	High Risk
13	CC-P25	Low Risk	CC-P25	Low Risk
14	CC-P28	High Risk	CC-P28	High Risk
15	CC-P29	High Risk	CC-P31	Low Risk
16	CC-P31	Low Risk	CC-P35	Low Risk
17	CC-P34	High Risk	CC-P36	Low Risk
18	CC-P35	High Risk	CC-P37	Low Risk
19	CC-P37	Low Risk	CC-P38	Low Risk
20	CC-P38	Low Risk	CC-P40	Low Risk
21	CC-P40	Low Risk	CC-P48	Low Risk
22	CC-P42	Low Risk	CC-P50	Low Risk
23	CC-P44	Low Risk	CC-P51	Low Risk
24	CC-P46	Low Risk	CC-P60	Low Risk
25	CC-P47	Low Risk	CC-P62	Low Risk
26	CC-P48	High Risk	CC-P66	Low Risk
27	CC-P50	Low Risk	CC-P71	High Risk
28	CC-P51	Low Risk	CC-P72	Low Risk
29	CC-P55	Low Risk	CC-P73	Low Risk
30	CC-P56	Low Risk		
31	CC-P60	Low Risk		
32	CC-P62	Low Risk		
33	CC-P66	Low Risk		
34	CC-P68	Low Risk		
35	CC-P70	High Risk		
36	CC-P71	High Risk		
37	CC-P72	High Risk		

4.5.5 Independence of 5-gene recurrence signature of tumor stage

This part of the study was focused on verifying if the recurrence predictions obtained on Ried et al data were statistically significant when validated separately in Stage II and Stage III patients. It was seen that the 5-gene signature could stratify the patients into low-risk and high-risk groups in Stage II and Stage III samples individually with distinct relapse-free survival. The patient subgroups were obtained based on the Random Tree model. The patients belonging to the low-risk group had higher survival probabilities than those belonging to the high-risk group. Based on the predictions from Cox model using the 5-gene signature, Kaplan-Meier plots were plotted in Stage II and Stage III samples separately. But the patient stratification was not statistically significant and the results were not reported. So it can be said that Random Tree model is the best model for predicting recurrence using the 5-gene signature. These results confirm that the 5-gene recurrence signature might be applicable to prognostic categorization for the clinical management of colon cancer.

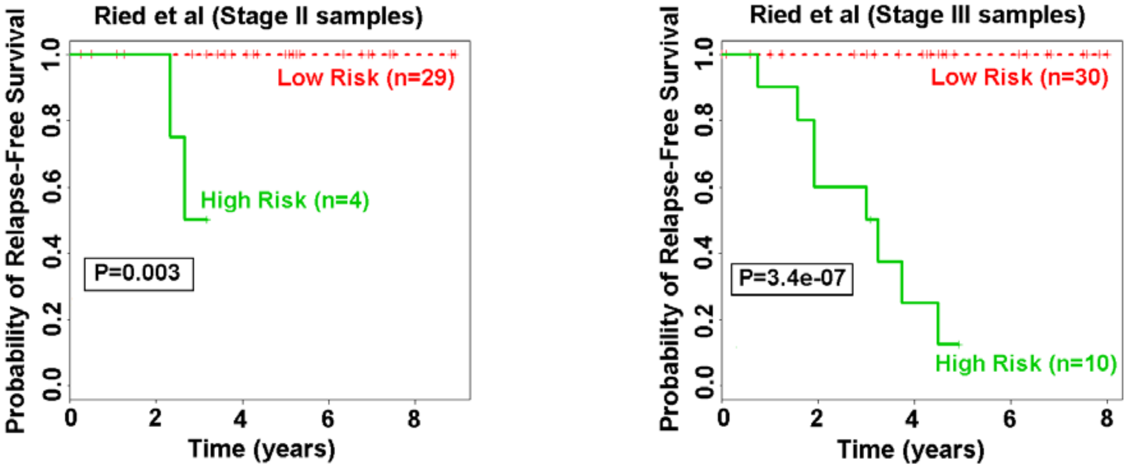

Figure 4.12 The 5-gene signature stratifies patients in Stage II tumors and Stage III tumors into distinct low-risk and high-risk groups for relapse-free survival using the random tree model.

4.5.6 External validation of the 5-gene signature on other colon cancer data

This part of the study sought to explore the extent to which the 5-gene signature could be used for prediction of lymph node status, recurrence, and drug response in publicly available independent datasets. More than 50 classifiers available in Weka software were tested using a leave-one-out cross validation technique on each of the independent datasets to find a suitable classification scheme for validation. Due to the different number of attributes (matching genes), sample sizes and prediction variables one specific scheme could not be used for validation on all the datasets. Different classifiers had to be employed on the validation datasets to get fair prediction accuracy. As far as possible the same set of classifiers were presented in the comparison tables of validation datasets to provide a fair evaluation of the performance. The exact same set of classifiers could not be compared over all the validation datasets due to poor performances of classifiers on some datasets and good performances on other datasets. The following sections discuss the validation results and comparisons of various classifiers on the independent datasets in detail.

4.5.6.1 Predicting lymph node status by leave-one-out cross validation on data from Koinuma et al. (*n=17*) PMID 16247484

The data from Koinuma et al (Affymetrix HG U133 B platform) consisted of 20 patient samples of which 3 patients were Duke's stage D. The Duke's stage D patients were not considered for validation. The search for matching genes with the 5-gene signature was done using the Affymetrix ids. There were 2 matching genes (Table 4.33). The data used for validation consisted of the expression of these 2 genes in the 17 patient samples. Weka software was used for validation and lymph node status (positive/negative) was predicted. Different classification schemes including *KStar*, Random Tree, Threshold selector, Multilayer perceptron, and AD Tree were applied to this dataset to find the best scheme. Table 4.34 shows the comparison between *KStar* and some of the classifiers used for validation on other datasets. *KStar* classifier performed better than the other classifiers. It had a sensitivity of 57.10%, a specificity of 70.00%, and an overall accuracy of 64.70%. Table 4.35 shows the confusion matrix for *KStar* classifier.

The difference in overall accuracy between *KStar* and other classifiers was not statistically significant due to the small sample size.

Table 4.33 Matching genes in Koinuma et al data.

GENE NAME	ID
HSPA14-heat shock 70kDa protein 14	H200018991
SLC25A5-solute carrier family 25	H200006643

Table 4.34 Comparison of accuracies obtained from different classifiers for predicting lymph node status using the 5-gene signature. The improved overall accuracy of the prediction with the *KStar* classifier compared with other methods was assessed by significance testing (*N* = 17).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
KStar	**57.10**	**70.00**	**63.55**	**64.70**	
Random Tree	57.10	60.00	58.55	58.82	< 0.37
Threshold selector	85.70	40.00	62.85	58.82	< 0.37
Multilayer perceptron	28.60	50.00	39.30	41.17	< 0.09
AD Tree	28.60	50.00	39.30	41.17	< 0.09

Table 4.35 Confusion matrix obtained from the *KStar* classifier for predicting lymph node status using the 5-gene signature.

Actual/Predicted	a (positive)	b (negative)
a (positive)	4	3
b (negative)	3	7

4.5.6.2 Predicting recurrence by leave-one-out cross validation on data from Barrier et al. (*n=18*) PMID 16091735

The data from Barrier et al (PMID 16091735) consisted of 22,283 genes and 18 patient samples. The search for matching genes with the 5-gene signature was done using the Affymetrix ids. There were 4 matching genes (Table 4.36). The data used for validation consisted of the expression of these 4 genes in the 18 patients. Weka software was used for validation and recurrence (yes/no) was predicted. Different classification schemes including AD Tree, Random Tree, Threshold selector, *KStar*, and Multilayer perceptron were applied to this dataset to find the best scheme. Table 4.37 shows the comparison between AD Tree and some of the classifiers used for validation on other datasets. AD Tree classifier performed better than the other classifiers. It had a sensitivity of 88.88%, a specificity of 88.88%, and an overall accuracy of 88.88%. Table 4.38 shows the confusion matrix for AD Tree classifier. The difference in overall accuracy between AD Tree and other classifiers was not statistically significant due to the small sample size.

Table 4.36 Matching genes in Barrier et al data.

GENE NAME	ID
TPD52L2-tumor protein D52-like2	H200013992
ZNF187-zinc finger protein 187 (ZNF187)	H200015602
HSPA14-heat shock 70kDa protein 14	H200018991
SLC25A5-solute carrier family 25	H200006643

Table 4.37 Comparison of accuracies obtained from different classifiers for predicting recurrence using the 5-gene signature. The improved overall accuracy of the prediction with the AD Tree classifier compared with other methods was assessed by significance testing (*N* = 18).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
AD Tree	**88.88**	**88.88**	**88.88**	**88.88**	
Random Tree	66.66	66.66	66.66	66.66	< 0.06
Threshold selector	77.80	66.70	72.25	72.22	< 0.10
KStar	77.80	55.60	66.70	66.66	< 0.06
Multilayer perceptron	77.80	77.80	77.80	77.77	< 0.18

Table 4.38 Confusion matrix obtained from the AD Tree classifier for predicting recurrence using the 5-gene signature.

Actual/Predicted	a (recurrence)	b (no recurrence)
a (recurrence)	8	1
b (no recurrence)	1	8

4.5.6.3 Predicting recurrence by leave-one-out cross validation on data from Barrier et al. (*n=50*) (PMID 16966692)

The data from Barrier et al (PMID 16966692) consisted of 22,283 genes and 50 patient samples. The search for matching genes with the 5-gene signature was done using the Affymetrix ids. There were 4 matching genes (Table 4.39). The data used for validation consisted of the expression of these 4 genes in the 50 patient samples. Weka software was used for validation and recurrence (yes/no) was predicted. Different classification schemes including Threshold selector, *KStar*, IB1, AD Tree, and Multilayer perceptron were applied to this dataset to find the best scheme. Table 4.40 shows the comparison between Threshold selector and some of the classifiers used for validation on other datasets. Threshold selector classifier performed better than the other classifiers. It had a sensitivity of 84.00%, a specificity of 68.00%, and an overall accuracy of 76.00%. Table 4.41 shows the confusion matrix for Threshold selector classifier. The difference in overall accuracy between Threshold selector and other classifiers was not statistically significant due to the small sample size.

Table 4.39 Matching genes in Barrier et al data.

GENE NAME	ID
TPD52L2-tumor protein D52-like2	H200013992
ZNF187-zinc finger protein 187 (ZNF187)	H200015602
HSPA14-heat shock 70kDa protein 14	H200018991
SLC25A5-solute carrier family 25	H200006643

Table 4.40 Comparison of accuracies obtained from different classifiers for predicting recurrence using the 5-gene signature. The improved overall accuracy of the prediction with the Threshold selector classifier compared with other methods was assessed by significance testing (*N* = 50).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Threshold selector	**84.00**	**68.00**	**76.00**	**76.00**	
KStar	80.00	68.00	74.00	74.00	< 0.40
IB1	76.00	72.00	74.00	74.00	< 0.40
AD Tree	76.00	64.00	70.00	70.00	< 0.24
Multilayer perceptron	60.00	68.00	64.00	64.00	< 0.09

Table 4.41 Confusion matrix obtained from the Threshold selector classifier for predicting recurrence using the 5-gene signature.

Actual/Predicted	a (no recurrence)	b (recurrence)
a (no recurrence)	17	8
b (recurrence)	4	21

4.5.6.4 Predicting recurrence by leave-one-out cross validation on data from Barrier et al. (*n=24*) (PMID 17043639)

The data from Barrier et al (PMID 17043639) consisted of 22,283 genes and 24 patient samples. The search for matching genes was done using the Affymetrix ids. There were 4 matching genes (Table 4.42). The data used for validation consisted of the expression of these 4 genes in the 24 patient samples. Weka software was used for validation and recurrence (yes/no) was predicted. Different classification schemes including Threshold selector, Logistic regression, LWL, Multilayer perceptron, and AD Tree were applied to this dataset to find the best scheme. Table 4.43 shows the comparison between Threshold selector and some of the classifiers used for validation on other datasets. Threshold selector classifier performed better than the other classifiers. It had a sensitivity of 50.00%, a specificity of 92.90%, and an overall accuracy of 75.00%. Table 4.44 shows the confusion matrix for Threshold selector classifier. The difference in overall accuracy between Threshold selector and other classifiers was not statistically significant due to the small sample size.

Table 4.42 Matching genes in Barrier et al data.

GENE NAME	ID
TPD52L2-tumor protein D52-like2	H200013992
ZNF187-zinc finger protein 187 (ZNF187)	H200015602
HSPA14-heat shock 70kDa protein 14	H200018991
SLC25A5-solute carrier family 25	H200006643

Table 4.43 Comparison of accuracies obtained from different classifiers for predicting recurrence using the 5-gene signature. The improved overall accuracy of the prediction with the Threshold selector classifier compared with other methods was assessed by significance testing (*N* = 24).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Threshold selector	**50.00**	**92.90**	**47.63**	**75.00**	
Logistic regression	60.00	78.60	69.30	70.83	< 0.38
LWL	70.00	71.40	70.70	70.83	< 0.38
Multilayer perceptron	60.00	78.60	69.30	70.83	< 0.38
AD Tree	70.00	64.30	67.15	66.66	< 0.27

Table 4.44 Confusion matrix obtained from the Threshold selector classifier for predicting recurrence using the 5-gene signature.

Actual/Predicted	a (no recurrence)	b (recurrence)
a (no recurrence)	13	1
b (recurrence)	5	5

4.5.6.5 Predicting the response of cell lines in NCI-60 (U133A GCRMA) data (*n*=*34*) by leave-one-out cross validation

This dataset[34] consisted of 21,225 genes and 60 cell lines (41). Our focus was on the sensitive and resistant cell lines, so cell lines with intermediate response were not considered for validation. A total of 34 cell lines (17 sensitive and the other 17 resistant to the drug 5-FU) were used in validation. The search for matching genes was done using the gene symbols. There were 4 matching genes (Table 4.45). The data used for validation consisted of the expression of these 4 genes in the 34 cell lines. Weka software was used for validation and the response (sensitive/resistant) for the drug 5-FU (fluorouracil) was predicted. Different classification schemes including Threshold selector, Multilayer perceptron, Random Tree, *KStar*, and AD Tree were applied to this dataset to find the best scheme. Table 4.46 shows the comparison between Threshold selector and some of the classifiers used for validation on other datasets. Threshold selector classifier performed better than the other classifiers. It had a sensitivity of 82.40%, a specificity of 64.70%, and an overall accuracy of 73.52%. Table 4.47 shows the confusion matrix for Threshold selector classifier. The difference in overall accuracy between Threshold selector and Multilayer perceptron ($p < 0.01$), *KStar* ($p < 0.04$), AD Tree ($p < 0.02$) was statistically significant.

Table 4.45 Matching genes in NCI-60 U133A data.

GENE NAME	ID
TPD52L2-tumor protein D52-like2	H200013992
ZNF187-zinc finger protein 187 (ZNF187)	H200015602
HSPA14-heat shock 70kDa protein 14	H200018991
SLC25A5-solute carrier family 25	H200006643

[34] http://discover.nci.nih.gov/cellminer/loadDownload.do

Table 4.46 Comparison of accuracies obtained from different classifiers for predicting drug response using the 5-gene signature. The improved overall accuracy of the prediction with the Threshold selector classifier compared with other methods was assessed by significance testing (*N* = 34).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Threshold selector	**82.40**	**64.70**	**73.55**	**73.52**	
Multilayer perceptron	52.90	35.30	44.10	44.11	< 0.01
Random Tree	58.80	52.90	55.85	55.88	< 0.07
KStar	64.70	41.20	52.95	52.94	< 0.04
AD Tree	41.20	52.90	47.05	47.05	< 0.02

Table 4.47 Confusion matrix obtained from the Threshold selector classifier for predicting drug response using the 5-gene signature.

Actual/Predicted	**a (sensitive)**	**b (resistant)**
a (sensitive)	14	3
b (resistant)	6	11

4.5.6.6 Predicting the response of cell lines in NCI-60 (U133B GCRMA) data (*n=34*) by leave-one-out cross validation

This dataset[35] consisted of 17910 genes and 60 cell lines (41). Our focus was on the sensitive and resistant cell lines, so cell lines with intermediate response were not considered for validation. A total of 34 cell lines (17 sensitive and the other 17 resistant to the drug 5-FU) were used in validation. The search for matching genes was done using the gene symbols. There was 1 matching gene (Table 4.48). The data used for validation consisted of the expression of this gene in the 34 cell lines. Weka software was used for validation and the response (sensitive/resistant) for the drug 5-FU (fluorouracil) was predicted. Different classification schemes including Threshold selector, AD Tree, Random Tree, *IB1*, and *KStar* were applied to this dataset to find the best scheme. Table 4.49 shows the comparison between Threshold selector and some of the classifiers used for validation on other datasets. Threshold selector classifier performed better than the other classifiers. It had a sensitivity of 82.35%, a specificity of 82.35%, and an

[35] http://discover.nci.nih.gov/cellminer/loadDownload.do

overall accuracy of 82.35%. Table 4.50 shows the confusion matrix for Threshold selector classifier. The difference in overall accuracy between Threshold selector and AD Tree ($p < 0.05$), Random Tree ($p < 0.01$), *IB1* ($p < 0.05$), *KStar* ($p < 0.01$) was statistically significant.

Table 4.48 Matching genes in NCI-60 U133B data.

GENE NAME	ID
HSPA14-heat shock 70kDa protein 14	H200018991

Table 4.49 Comparison of accuracies obtained from different classifiers for predicting drug response using the 5-gene signature. The improved overall accuracy of the prediction with the Threshold selector classifier compared with other methods was assessed by significance testing (*N* = 34).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Threshold selector	**82.35**	**82.35**	**82.35**	**82.35**	
AD Tree	64.70	64.70	64.70	64.70	< 0.05
Random Tree	52.90	58.80	55.85	55.88	< 0.01
IB1	52.90	47.10	50.00	50.00	< 0.01
KStar	58.80	29.40	44.10	44.11	< 0.01

Table 4.50 Confusion matrix obtained from the Threshold selector classifier for predicting drug response using the 5-gene signature.

Actual/Predicted	a (sensitive)	b (resistant)
a (sensitive)	14	3
b (resistant)	3	14

4.5.7 Summary of validation results of 5-gene signature

Table 4.51 shows the details of different validation datasets, predicted variables, classifiers used and different accuracies obtained using the 5-gene signature. For each dataset the classifier with the highest overall accuracy was reported.

Table 4.51 Summary of validation results of 3-gene signature on Ried et al data, independent colon cancer datasets and NCI 60 data.

Dataset	Classifier	Predicted variable	Sensitivity (%)	Specificity (%)	(Sensitivity + Specificity)/2 (%)	Overall accuracy (%)
Ried et al training set (*n*=36) PMID 17210682	Random Tree	Recurrence	70.00	88.46	79.23	83.33
Ried et al training set (*n*=36) PMID 17210682	Cox model	Recurrence	70.00	80.80	75.40	77.80
Koinuma et al (*n*=17) PMID 16247484	KStar	Lymph node status	57.10	70.00	63.55	64.70
Barrier et al (*n*=18) PMID 16091735	AD Tree	Recurrence	88.88	88.88	88.88	88.88
Barrier et al (*n*=50) PMID 16966692	Threshold selector	Recurrence	84.00	68.00	76.00	76.00
Barrier et al (n=24) PMID 17043639	Threshold selector	Recurrence	50.00	92.90	47.63	75.00
NCI 60 U133A (*n*=34)	Threshold selector	Drug response (5-FU)	82.40	64.70	73.55	73.52
NCI 60 U133B (*n*=34)	Threshold selector	Drug response (5-FU)	82.35	82.35	82.35	82.35

4.6 Comparison of 3-gene and 5-gene signatures

This part of the study discusses the 3-gene and 5-gene recurrence signatures and compares them. The gene *SLC25A5* was common in both the gene signatures. Based on the prediction accuracies obtained from the independent validation datasets, it can be seen that the 3-gene signature performs better than the 5-gene signature. But the patient stratification in Stage II and Stage II tumor samples, by both the gene signatures were statistically significant. It can be concluded that both the 3-gene and 5-gene signatures could be used to predict recurrence and identify patients at high-risk of recurrence. Table 4.52 shows the comparison between 3-gene signature and 5-gene signature in detail. The difference in the prediction accuracies obtained from 3-gene and 5-gene signature on NCI-60 U133A data was statistically significant, whereas the results on other datasets were not significant.

Table 4.52 Comparison of prediction accuracies obtained from 3-gene and 5-gene signatures on independent datasets. The improved overall accuracy of the prediction with the 3-gene signature compared with the 5-gene signature was assessed by significance testing.

Dataset		**Ried et al training dataset (*n=36*)**	**Ried et al training dataset (*n=36*)**	**Koinuma et al data (*n=17*)**	**Barrier et al data (*n=18*)**	**Barrier et al data (*n=50*)**	**NCI-60 U133A data (*n=34*)**	**NCI-60 U133B data (*n=34*)**
Predicted variable		Recurrence	Recurrence	Lymph node status	Recurrence	Recurrence	Drug response	Drug response
3-gene signature	**Classifier**	LWL	Cox model	Multilayer perceptron	Threshold selector	IB1	Threshold selector	Threshold selector
	Sensitivity (%)	80.00	80.00	71.40	88.90	76.00	94.10	88.23
	Specificity (%)	96.20	96.10	90.00	77.80	80.00	88.20	88.23
	Overall accuracy (%)	91.66	91.70	82.35	83.33	78.00	91.17	88.23
5-gene signature	**Classifier**	Random Tree	Cox model	KStar	AD Tree	Threshold selector	Threshold selector	Threshold selector
	Sensitivity (%)	70.00	70.00	57.10	88.88	84.00	82.40	82.35
	Specificity (%)	88.46	80.80	70.00	88.88	68.00	64.70	82.35
	Overall accuracy (%)	83.33	77.80	64.70	88.88	76.00	73.52	82.35
***P*-value**		< 0.15	< 0.06	< 0.13	< 0.31	< 0.41	< 0.03	< 0.25

4.7 Summary

In this chapter, we described how the recurrence gene signatures were identified. A combinatorial scheme was utilized for feature selection. Firstly, variable selection using random forests was applied on the preprocessed data to identify gene subsets, and secondly, InfoGain attribute selection technique was applied to reduce the dimensionality of the gene signatures without decreasing the predictive power. Two prediction models were built independently with the 3-gene and the 5-gene signatures using classifiers in Weka software to predict the risk stage of the patients in the testing set (patients whose recurrence status is currently unknown). The subgroups of patients without recurrence and survival time more than 5 years, and the patients having recurrence within 5 years after surgery obtained from the Cox model had a sensitivity of 80.0% and a specificity of 96.1%, using the 3-gene signature. Using the 5-gene signature, the sensitivity was 70.0% and the specificity was 80.8%. The Kaplan-Meier plots for the 3-gene signature and the 5-gene signature on Ried et al data obtained based on the Cox model stratified patients into distinct low-risk and high-risk groups. Both the gene signatures were cross validated on independent colon cancer data sets. The drug response of 5-FU (fluorouracil) on the NCI-60 cell line data was predicted. To confirm the prognostic applicability of the recurrence gene signatures, Kaplan Meier curves were plotted separately for Stage II and Stage III patients based on the predicted subgroups. The stratification was statistically significant (log-rank tests, p<0.05) for the 3-gene and 5-gene signatures. This confirms that it is feasible to predict recurrence in the Stage II and Stage III tumors with the 3-gene and 5-gene signatures.

Chapter 5

Validation of the identified gene signatures on rectal cancer data

5.1 Introduction

As colon cancer and rectal cancer are anatomically related, this part of the study sought to explore whether the identified colon cancer gene signatures could predict lymph node metastasis and generate significant patient stratification into low-risk and high-risk groups on rectal cancer data. The rectal cancer data was obtained from Ried et al (*n=29*) (PMID 16397240) (32). The 29 patients included in this study were all participants in a multicenter, randomized prospective phase III clinical trial treated at the Department of General Surgery, University Medical Center Gottingen, Germany. This data set of 29 carcinomas and 20 mucosa biopsies includes 12 patient-matched pairs of biopsies from tumor and normal mucosa. The lymph node status, chemoradiotherapy response, disease-free survival, and overall survival information was available for all the patients. All the patients received a dose of 50.4Gy of radiation accompanied by FU (Fluorouracil). The following sections describe the validation results of the 9-gene lymph node status signature, 3-gene and 5-gene recurrence signatures on rectal cancer data including time-dependent ROC and Kaplan-Meier analyses.

5.2 Validation results of the 9-gene signature on rectal cancer data

5.2.1 Predicting lymph node status by leave-one-out cross validation on cDNA 1 files

The cDNA 1 data files from Ried et al (PMID 16397240) consisted of 23 patient samples. The search for matching genes was done using the gene symbols. There were 3 matching genes (Table 5.1). The data used for validation consisted of the expression of these 3 genes in the 23 patient samples. Weka software was used for validation and lymph node status (negative/positive) was predicted. Different classification schemes including *AdaboostM1*, Multiboost AB, Random Tree, *IB1*, and Multilayer perceptron were applied to this dataset to find the best scheme. Table 5.2 shows the comparison between *AdaboostM1* and some of the classifiers used for validation on other datasets. *AdaboostM1* classifier performed better than the other classifiers. It had a sensitivity of 87.50%, a specificity of 57.10%, and an overall accuracy of 78.26%. Table 5.3 shows the confusion matrix for *AdaboostM1* classifier. The difference in overall accuracy between *AdaboostM1* and other classifiers was not statistically significant.

Table 5.1 Matching genes in Ried et al rectal cancer data.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
PLXNB2-plexin B2, mRNA	H200000861
ITGB1-integrin,beta1 (fibronectin)	H200021334

Table 5.2 Comparison of accuracies obtained from different classifiers for predicting lymph node status using the 9-gene signature. The improved overall accuracy of the prediction with the *AdaboostM1* classifier compared with other methods was assessed by significance testing (*N* = 23).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity +Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
AdaboostM1	**87.50**	**57.10**	**72.30**	**78.26**	
Multiboost AB	81.30	28.60	54.95	65.21	<0.17
Random Tree	81.30	28.60	54.95	65.21	<0.17
IB1	87.50	42.90	65.20	73.91	<0.35
Multilayer perceptron	81.30	28.60	54.95	65.21	<0.17

Table 5.3 Confusion matrix obtained from the *AdaboostM1* classifier for predicting lymph node status using the 9-gene signature.

Actual/Predicted	a (node negative)	b (node positive)
a (node negative)	4	3
b (node positive)	2	14

5.2.2 Predicting lymph node status by leave-one-out cross validation on cDNA2 files

The cDNA 2 data files from Ried et al (PMID 16397240) consisted of 23 patient samples. The search for matching genes was done using the gene symbols. There were 3 matching genes (Table 5.4). The data used for validation consisted of the expression of these 3 genes in the 23 patient samples. Weka software was used for validation and lymph node status (negative/positive) was predicted. Different classification schemes including *AdaboostM1*, Multiboost AB, Random Tree, IB1, and *JRip* were applied to this dataset to find the best scheme. Table 5.5 shows the comparison between *AdaboostM1* and some of the classifiers used for validation on other datasets. *AdaboostM1* classifier performed better than the other classifiers. It had a sensitivity of 93.80%, a specificity of 42.90%, and an overall accuracy of 78.26%. Table 5.6 shows the confusion matrix for *AdaboostM1* classifier. The difference in overall accuracy between *AdaboostM1* and other classifiers was not statistically significant due to the small sample size.

Table 5.4 Matching genes in Ried et al rectal cancer data.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
PLXNB2-plexin B2, mRNA	H200000861
ITGB1-integrin,beta1 (fibronectin)	H200021334

Table 5.5 Comparison of accuracies obtained from different classifiers for predicting lymph node status using the 9-gene signature. The improved overall accuracy of the prediction with the *AdaboostM1* classifier compared with other methods was assessed by significance testing (*N* = 23).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
AdaboostM1	**93.80**	**42.90**	**68.35**	**78.26**	
Multiboost AB	75.00	14.30	44.65	56.52	<0.06
Random Tree	68.80	28.60	48.70	56.52	<0.06
IB1	68.80	42.90	55.85	60.86	<0.11
JRip	93.80	28.60	61.20	73.91	<0.37

Table 5.6 Confusion matrix obtained from the *AdaboostM1* classifier for predicting lymph node status using the 9-gene signature.

Actual/Predicted	a (node negative)	b (node positive)
a (node negative)	3	4
b (node positive)	1	15

5.2.3 Predicting lymph node status by leave-one-out cross validation on tumor biopsies 1 files

The tumor biopsies 1 data files from Ried et al (PMID 16397240) consisted of 17 patient samples. The search for matching genes was done using the gene symbols. There were 9 matching genes (Table 5.7).

The data used for validation consisted of the expression of these 9 genes in the 17 patient samples. Weka software was used for validation and lymph node status (negative/positive) was predicted. Different classification schemes including Decision stump, Multilayer perceptron, Random Tree, *AdaboostM1*, and *IB1* were applied to this dataset to find the best scheme. Table 5.8 shows the comparison between Decision stump and some of the classifiers used for validation on other datasets. Decision stump classifier performed better than the other classifiers. It had a sensitivity of 75.00%, a specificity of 80.00%, and an overall accuracy of 76.47%. Table 5.9 shows the confusion matrix for Decision stump classifier. The difference in overall accuracy between Decision stump and Multilayer perceptron ($p < 0.04$), Random Tree ($p < 0.04$) was statistically significant.

Table 5.7 Matching genes in Ried et al rectal cancer data.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
IFRG28-28kD interferon responsive pro	H200004627
PLXNB2-plexin B2, mRNA	H200000861
DC50-hypothetical protein DC50	H200019106
FLJ11078-hypothetical protein FLJ1107	H200016227
MGC16044-hypothetical protein MGC1604	H200020589
RNF6-ring finger protein (C3H2C3 type)	H200004174
POU6F2-POU domain, class 6,transcript	H200015474
ITGB1-integrin,beta1 (fibronectin)	H200021334

Table 5.8 Comparison of accuracies obtained from different classifiers for predicting lymph node status using the 9-gene signature. The improved overall accuracy of the prediction with the Decision stump classifier compared with other methods was assessed by significance testing (*N* = 17).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Decision stump	**75.00**	**80.00**	**77.50**	**76.47**	
Multilayer perceptron	58.30	20.00	39.15	47.05	<0.04
Random Tree	58.30	20.00	39.15	47.05	<0.04
AdaboostM1	66.70	20.00	43.33	52.94	<0.08
IB1	58.30	40.00	49.15	52.94	<0.08

Table 5.9 Confusion matrix obtained from the Decision stump classifier for predicting lymph node status using the 9-gene signature.

Actual/Predicted	a (node positive)	b (node negative)
a (node positive)	9	3
b (node negative)	1	4

5.2.4 Predicting lymph node status by leave-one-out cross validation on tumor biopsies 2 files

The tumor biopsies 2 data files from Ried et al (PMID 16397240) consisted of 17 patient samples. The search for matching genes was done using the gene symbols. There were 9 matching genes (Table 5.10). The data used for validation consisted of the expression of these 9 genes in the 17 patient samples. Weka software was used for validation and lymph node status (negative/positive) was predicted. Different classification schemes including *J48*, Random Tree, Adaboost M1, Multiboost AB, and Multilayer perceptron were applied to this dataset to find the best scheme. Table 5.11 shows the comparison between *J48* and some of the classifiers used for validation on other datasets. *J48* classifier performed better than the other classifiers. It had a sensitivity of 75.00%, a specificity of 60.00%, and an overall accuracy of 70.58%. Table 5.12 shows the confusion matrix for *J48* classifier. The difference in overall accuracy between *J48* and other classifiers was not statistically significant due to the small sample size.

Table 5.10 Matching genes in Ried et al rectal cancer data.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
IFRG28-28kD interferon responsive pro	H200004627
PLXNB2-plexin B2, mRNA	H200000861
DC50-hypothetical protein DC50	H200019106
FLJ11078-hypothetical protein FLJ1107	H200016227
MGC16044-hypothetical protein MGC1604	H200020589
RNF6-ring finger protein (C3H2C3 type)	H200004174
POU6F2-POU domain, class 6,transcript	H200015474
ITGB1-integrin,beta1 (fibronectin)	H200021334

Table 5.11 Comparison of accuracies obtained from different classifiers for predicting lymph node status using the 9-gene signature. The improved overall accuracy of the prediction with the *J48* classifier compared with other methods was assessed by significance testing (*N* = 17).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
J48	**75.00**	**60.00**	**67.50**	**70.58**	
Random Tree	75.00	40.00	57.50	64.70	<0.36
Adaboost M1	66.70	40.00	53.35	58.82	<0.24
Multiboost AB	75.00	20.00	47.50	58.82	<0.24
Multilayer perceptron	66.70	40.00	53.35	58.82	<0.24

Table 5.12 Confusion matrix obtained from the *J48* classifier for predicting lymph node status using the 9-gene signature.

Actual/Predicted	**a (node positive)**	**b (node negative)**
a (node positive)	9	3
b (node negative)	2	3

5.2.5 Predicting chemoradiotherapy response by leave-one-out cross validation on cDNA 1 files

The cDNA 1 data files from Ried et al (PMID 16397240) consisted of 23 patient samples. The search for matching genes was done using the gene symbols. There were 3 matching genes (Table 5.13). The data used for validation consisted of the expression of these 3 genes in the 23 patient samples. Weka software was used for validation and chemoradiotherapy response (yes/no) was predicted. Different classification schemes including *JRip*, *J48*, *AdaboostM1*, Random Tree, and Multilayer perceptron were applied to this dataset to find the best scheme. Table 5.14 shows the comparison between *JRip* and some of the classifiers used for validation on other datasets. *JRip* classifier performed better than the other classifiers. It had a sensitivity of 55.60%, a specificity of 85.70%, and an overall accuracy of 73.91%. Table 5.15 shows the confusion matrix for *JRip* classifier. The difference in overall accuracy between *JRip* and *J48* (p < 0.04) was statistically significant.

Table 5.13 Matching genes in Ried et al rectal cancer data.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
PLXNB2-plexin B2, mRNA	H200000861
ITGB1-integrin,beta1 (fibronectin)	H200021334

Table 5.14 Comparison of accuracies obtained from different classifiers for predicting chemoradiotherapy response using the 9-gene signature. The improved overall accuracy of the prediction with the *JRip* classifier compared with other methods was assessed by significance testing (*N* = 23).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
JRip	**55.60**	**85.70**	**70.65**	**73.91**	
J48	33.30	57.10	45.20	47.82	<0.04
AdaboostM1	44.40	78.60	61.50	65.21	<0.27
Random Tree	44.40	71.40	57.90	60.86	<0.18
Multilayer perceptron	55.60	64.30	59.95	60.86	<0.18

Table 5.15 Confusion matrix obtained from the *JRip* classifier for predicting response using the 9-gene signature.

Actual/Predicted	**a (response)**	**b (no response)**
a (response)	5	4
b (no response)	2	12

5.2.6 Predicting chemoradiotherapy response by leave-one-out cross validation on cDNA2 files

The cDNA 2 data files from Ried et al (PMID 16397240) consisted of 23 patient samples. The search for matching genes was done using the gene symbols. There were 3 matching genes (Table 5.16). The data used for validation consisted of the expression of these 3 genes in the 23 patient samples. Weka software was used for validation and chemoradiotherapy response (yes/no) was predicted. Different classification schemes including *JRip*, *AdaboostM1*, *IB1*, AD Tree, and Multiboost AB were applied to this dataset to find the best scheme. Table 5.17 shows the comparison between *JRip* and some of the classifiers used for validation on other datasets. *JRip* classifier performed better than the other classifiers. It had a sensitivity of 77.80%, a specificity of 85.70%, and an overall accuracy of 82.60%. Table 5.18 shows the confusion matrix for *JRip* classifier. The difference in overall accuracy between *JRip* and other classifiers was not statistically significant due to the small sample size.

Table 5.16 Matching genes in Ried et al rectal cancer data.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
PLXNB2-plexin B2, mRNA	H200000861
ITGB1-integrin,beta1 (fibronectin)	H200021334

Table 5.17 Comparison of accuracies obtained from different classifiers for predicting chemoradiotherapy response using the 9-gene signature. The improved overall accuracy of the prediction with the *JRip* classifier compared with other methods was assessed by significance testing (N = 23).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
JRip	**77.80**	**85.70**	**81.75**	**82.60**	
AdaboostM1	66.70	78.60	72.65	73.91	<0.24
IB1	44.40	71.40	57.90	60.86	<0.06
AD Tree	55.60	78.60	67.10	69.56	<0.15
Multiboost AB	55.60	71.40	63.50	65.21	<0.10

Table 5.18 Confusion matrix obtained from the *JRip* classifier for predicting response using the 9-gene signature.

Actual/Predicted	a (response)	b (no response)
a (response)	7	2
b (no response)	2	12

5.2.7 Predicting chemoradiotherapy response by leave-one-out cross validation on tumor biopsies 1 files

The tumor biopsies 1 data files from Ried et al (PMID 16397240) consisted of 17 patient samples. The search for matching genes was done using the gene symbols. There were 9 matching genes (Table 5.19). The data used for validation consisted of the expression of these 9 genes in the 17 patient samples. Weka software was used for validation and chemoradiotherapy response (yes/no) was predicted. Different classification schemes including Random committee, Multiboost AB, *IB1*, Multilayer perceptron, and

KStar were applied to this dataset to find the best scheme. Table 5.20 shows the comparison between Random committee and some of the classifiers used for validation on other datasets. Random committee classifier performed better than the other classifiers. It had a sensitivity of 90.00%, a specificity of 28.60%, and an overall accuracy of 64.70%. Table 5.21 shows the confusion matrix for Random committee classifier. The difference in overall accuracy between Random committee and other classifiers was not statistically significant due to the small sample size.

Table 5.19 Matching genes in Ried et al rectal cancer data.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
IFRG28-28kD interferon responsive pro	H200004627
PLXNB2-plexin B2, mRNA	H200000861
DC50-hypothetical protein DC50	H200019106
FLJ11078-hypothetical protein FLJ1107	H200016227
MGC16044-hypothetical protein MGC1604	H200020589
RNF6-ring finger protein (C3H2C3 type)	H200004174
POU6F2-POU domain, class 6,transcript	H200015474
ITGB1-integrin,beta1 (fibronectin)	H200021334

Table 5.20 Comparison of accuracies obtained from different classifiers for predicting chemoradiotherapy response using the 9-gene signature. The improved overall accuracy of the prediction with the Random committee classifier compared with other methods was assessed by significance testing (*N* = 17).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Random committee	**90.00**	**28.60**	**59.30**	**64.70**	
Multiboost AB	80.00	14.30	47.15	52.94	<0.25
IB1	70.00	42.90	56.45	58.82	<0.37
Multilayer perceptron	50.00	28.60	39.30	41.17	<0.09
KStar	70.00	42.90	56.45	58.82	<0.37

Table 5.21 Confusion matrix obtained from the Random committee classifier for predicting response using the 9-gene signature.

Actual/Predicted	**a (response)**	**b (no response)**
a (response)	9	1
b (no response)	5	2

5.2.8 Predicting chemoradiotherapy response by leave-one-out cross validation on tumor biopsies 2 files

The tumor biopsies 2 data files from Ried et al (PMID 16397240) consisted of 17 patient samples. The search for matching genes was done using the gene symbols. There were 9 matching genes (Table 5.22). The data used for validation consisted of the expression of these 9 genes in the 17 patient samples. Weka software was used for validation and chemoradiotherapy response (yes/no) was predicted. Different classification schemes including Logistic regression, *IB1*, *AdaboostM1*, Multilayer perceptron, and Threshold selector were applied to this dataset to find the best scheme. Table 5.23 shows the comparison between Logistic regression and some of the classifiers used for validation on other datasets. Logistic regression classifier performed better than the other classifiers. It had a sensitivity of 80.00%, a specificity of 71.40%, and an overall accuracy of 76.47%. Table 5.24 shows the confusion matrix for Logistic regression classifier. The difference in overall accuracy between Logistic regression and *AdaboostM1* (p < 0.04) was statistically significant.

Table 5.22 Matching genes in Ried et al rectal cancer data.

GENE NAME	ID
SNRPD3-small nuclear ribonucleoprotein	H200000411
IFRG28-28kD interferon responsive pro	H200004627
PLXNB2-plexin B2, mRNA	H200000861
DC50-hypothetical protein DC50	H200019106
FLJ11078-hypothetical protein FLJ1107	H200016227
MGC16044-hypothetical protein MGC1604	H200020589
RNF6-ring finger protein (C3H2C3 type)	H200004174
POU6F2-POU domain, class 6,transcript	H200015474
ITGB1-integrin,beta1 (fibronectin)	H200021334

Table 5.23 Comparison of accuracies obtained from different classifiers for predicting chemoradiotherapy response using the 9-gene signature. The improved overall accuracy of the prediction with the Logistic regression classifier compared with other methods was assessed by significance testing (*N* = 17).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Logistic regression	**80.00**	**71.40**	**75.70**	**76.47**	
IB1	80.00	42.90	61.45	64.70	<0.23
AdaboostM1	60.00	28.60	44.30	47.05	<0.04
Multilayer perceptron	70.00	57.10	63.55	64.70	<0.23
Threshold selector	40.00	85.70	62.85	58.82	<0.13

Table 5.24 Confusion matrix obtained from the Logistic regression classifier for predicting response using the 9-gene signature.

Actual/Predicted	**a (response)**	**b (no response)**
a (response)	8	2
b (no response)	2	5

5.2.9 Time-dependent ROC analyses on rectal cancer data from Ried et al (*n=23*) using the 9-gene lymph node status signature

To explore whether the 9-gene lymph node status signature could predict patient disease-free survival and overall survival, the survival and status information along with the expression data of the matching genes are used for getting the time-dependent ROC curves. There were 3 matching genes with the 9-gene signature. The expression data of these 3 genes in the 23 patient samples along with the survival information was used to plot the time-dependent ROC curves. The accuracy of 5-year disease-free survival prediction is 0.72 and 5-year overall survival prediction is 0.76, as represented by AUC for cDNA 1 data files. The accuracy of 5-year disease-free survival prediction is 0.79 and 5-year overall survival prediction is 0.75, as represented by AUC for cDNA 2 data files.

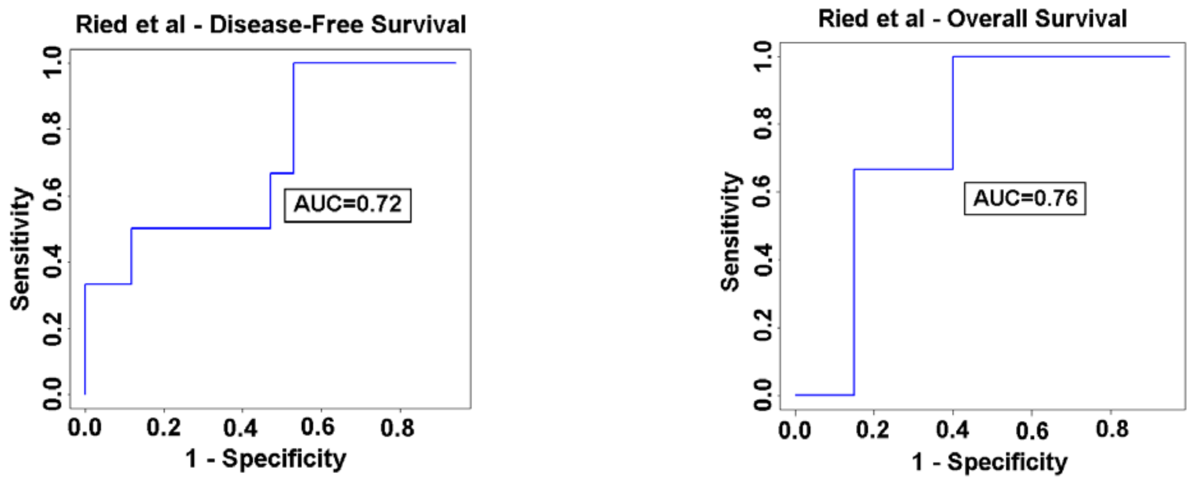

Figure 5.1 Time-dependent ROC plots on rectal cancer data (*n=23*) for disease-free survival and overall survival using the 9-gene signature in cDNA1 data files.

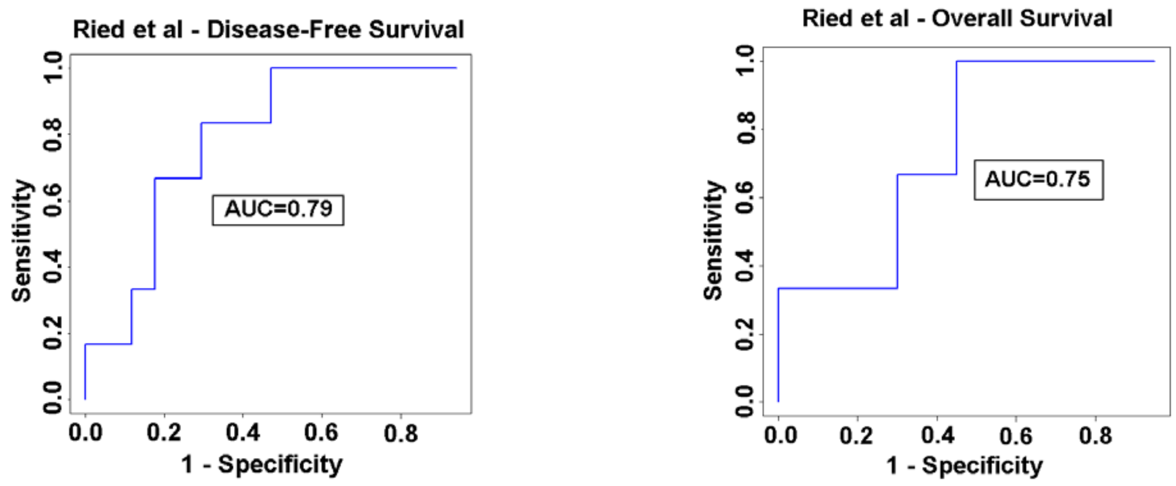

Figure 5.2 Time-dependent ROC plots on rectal cancer data (*n=23*) for disease-free survival and overall survival using the 9-gene signature in cDNA 2 data files.

5.2.10 Kaplan-Meier analyses on Ried et al rectal cancer data (*n=23*) using the 9-gene lymph node status signature

The cDNA 1 files in rectal cancer data were checked for matching genes with the 9-gene signature. There were 3 matching genes. The Cox model based on the expression of these 3 genes was used to get recurrence risk scores for the 23 patients. The choices for choosing a cut-off value for patient stratification are the peak value from histogram, mean risk score or median risk score. In this analysis, the median risk score was chosen as cut-off as it resulted in best patient stratification. Cut-off values of 0.41 and 0.09 were chosen for relapse-free survival and overall survival in cDNA 1 files, respectively. The *pamr* package in R was used to plot the Kaplan-Meier curves. The low-risk and high-risk groups, had distinct relapse-free survival (*p = 0.014, n=23,* log-rank tests) and distinct overall survival (*p = 0.043, n=23,* log-rank tests), respectively for the data in cDNA1 files. Table 5.25 shows the different parameters obtained from the Cox model using the 9-gene signature for disease-free survival and overall survival in cDNA 1 data files.

Table 5.25 Different parameters obtained from Cox model using the 9-gene signature for disease-free survival and overall survival in cDNA 1 data files.

Disease-free survival						**Overall survival**				
Gene Symbol	**coef**	**exp (coef)**	**se (coef)**	**z-score**	***p-value***	**Coef**	**exp (coef)**	**se (coef)**	**z-score**	***p-value***
SNRPD3	-0.541	0.582	0.548	-0.987	0.32	0.476	1.61	0.84	0.567	0.57
PLXNB2	1.955	7.070	1.363	1.435	0.15	1.335	3.80	1.91	0.699	0.48
ITGB1	0.009	1.010	0.758	0.013	0.99	0.217	1.24	1.32	0.164	0.87

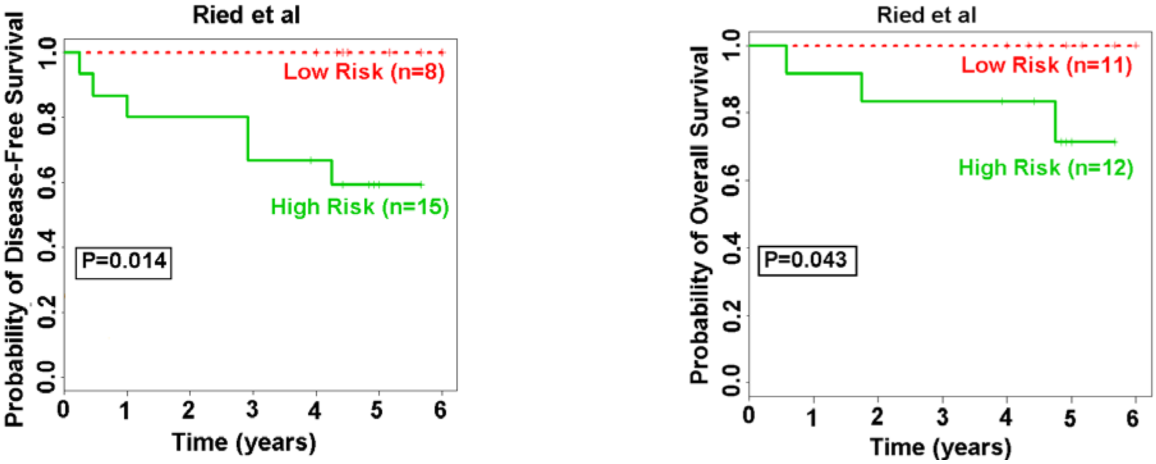

Figure 5.3 Kaplan-Meier plots on rectal cancer data (*n=23*) for disease-free survival and overall survival using the 9-gene signature in cDNA 1 files.

The cDNA 2 files in rectal cancer data were checked for matching genes with the 9-gene signature. There were 3 matching genes. The Cox model based on the expression of these 3 genes was used to get recurrence risk scores for the 23 patients. The choices for choosing a cut-off value for patient stratification are the peak value from histogram, mean risk score or median risk score. In this analysis, the median risk score was chosen as cut-off as it resulted in best patient stratification. Cut-off values 0.27 and -0.48 were chosen for relapse-free survival and overall survival in cDNA 2 files, respectively. The *pamr* package in R was used to plot the Kaplan-Meier curves. The low-risk and high-risk groups, had distinct relapse-free survival (*p = 0.041*, *n=23*, log-rank tests) and distinct overall survival (*p = 0.0436*, *n=23*, log-rank tests), respectively for the data in cDNA 2 files. Table 5.26 shows the different parameters obtained from the Cox model using the 9-gene signature for disease-free survival and overall survival in cDNA 2 data files.

Table 5.26 Different parameters obtained from Cox model using the 9-gene signature for disease-free survival and overall survival in cDNA 2 files.

	Disease-free survival					Overall survival				
Gene Symbol	**coef**	**exp (coef)**	**se (coef)**	**z-score**	***p-value***	**coef**	**exp (coef)**	**se (coef)**	**z-score**	***p-value***
SNRPD3	-0.122	0.885	0.671	-0.182	0.86	0.876	2.401	0.89	0.984	0.32
PLXNB2	1.262	3.531	1.247	1.012	0.31	-0.973	0.378	2.11	-0.461	0.65
ITGB1	-0.823	0.439	0.882	-0.933	0.35	-1.844	0.158	1.67	-1.104	0.27

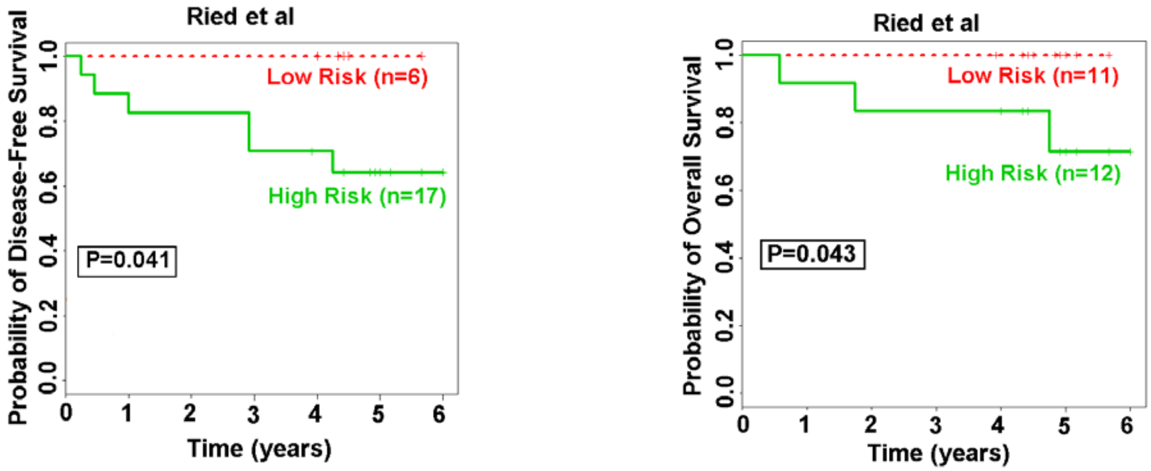

Figure 5.4 Kaplan-Meier plots on rectal cancer data (*n=23*) for disease-free survival and overall survival using the 9-gene signature in cDNA 2 data files.

5.2.11 Summary of validation results of 9-gene signature on rectal cancer data

Table 5.27 shows the details of validation results on rectal cancer data in different groups of files. For each dataset the classifier with the highest overall accuracy was reported. The time-dependent ROC and Kaplan-Meier analyses on tumor biopsies data were not reported as they were not significant.

Table 5.27 Summary of validation results of 9-gene signature on rectal cancer data.

Rectal cancer data	Classifier	Predicted variable	Sensitivity (%)	Specificity (%)	(Sensitivity + Specificity)/2 (%)	Overall accuracy (%)
cDNA1 files	AdaboostM1	Lymph node status	87.50	57.10	72.30	78.26
cDNA2 files	AdaboostM1	Lymph node status	93.80	42.90	68.35	78.26
Tumor biopsies 1 files	Decision stump	Lymph node status	75.00	80.00	77.50	76.47
Tumor biopsies 2 files	J48	Lymph node status	75.00	60.00	67.50	70.58
cDNA1 files	JRip	Chemoradiotherapy response	55.60	87.50	70.65	73.91
cDNA2 files	JRip	Chemoradiotherapy response	77.80	85.70	81.75	82.60
Tumor biopsies 1 files	Random Committee	Chemoradiotherapy response	90.00	28.60	59.30	64.70
Tumor biopsies 2 files	Logistic	Chemoradiotherapy response	80.00	71.40	75.70	76.47

5.3 Validation results of the 3-gene signature on rectal cancer data

5.3.1 Predicting lymph node status in tumor biopsies 2 files

The tumor biopsies 2 data files from Ried et al (PMID 16397240) consisted of 17 patient samples. The search for matching genes was done using the gene symbols. There were 2 matching genes (Table 5.28). The data used for validation consisted of the expression of these 2 genes in the 17 patient samples. Weka software was used for validation and lymph node status (positive/negative) was predicted. Different classification schemes including *KStar*, Logistic regression, AD Tree, *AdaboostM1*, and Threshold selector were applied to this dataset to find the best scheme. Table 5.29 shows the comparison between *KStar* and some of the classifiers used for validation on other datasets. *KStar* classifier performed better than the other classifiers. It had a sensitivity of 83.33%, a specificity of 60.00%, and an overall accuracy of 76.47%. Table 5.30 shows the confusion matrix for *KStar* classifier. The difference in overall accuracy between *KStar* and other classifiers was not statistically significant due to the small sample size.

Table 5.28 Matching genes in Ried et al rectal cancer data.

GENE NAME	ID
E2F2-E2F transcription factor 2 (E2F2)	H200012309
SLC25A5-solute carrier family 25	H200006643

Table 5.29 Comparison of accuracies obtained from different classifiers for predicting lymph node status using the 3-gene signature. The improved overall accuracy of the prediction with the *KStar* classifier compared with other methods was assessed by significance testing (*N* = 17).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity +Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
KStar	**83.33**	**60.00**	**71.66**	**76.47**	
Logistic regression	83.30	40.00	61.65	70.58	<0.35
AD Tree	66.70	60.00	63.35	64.70	<0.23
AdaboostM1	58.30	60.00	59.15	58.82	<0.14
Threshold selector	41.70	80.00	60.85	52.94	<0.08

Table 5.30 Confusion matrix obtained from the *KStar* classifier for predicting lymph node status using the 3-gene signature.

Actual/Predicted	**a (node positive)**	**b (node negative)**
a (node positive)	10	2
b (node negative)	2	3

5.3.2 Predicting chemoradiotherapy response in tumor biopsies 2 files

The tumor biopsies 2 data files from Ried et al (PMID 16397240) consisted of 17 patient samples. The search for matching genes was done using the gene symbols. There were 2 matching genes (Table 5.31). The data used for validation consisted of the expression of these 2 genes in the 17 patient samples. Weka software was used for validation and chemoradiotherapy response (yes/no) was predicted. Different classification schemes including Decision stump, Multiboost AB, *AdaboostM1*, Logistic regression, and AD Tree were applied to this dataset to find the best scheme. Table 5.32 shows the comparison between Decision stump and some of the classifiers used for validation on other datasets. Decision stump classifier performed better than the other classifiers. It had a sensitivity of 90.00%, a specificity of 57.10%, and an overall accuracy of 76.47%. Table 5.33 shows the confusion matrix for Decision stump classifier. The difference in overall accuracy between Decision stump and AD Tree ($p < 0.04$) was statistically significant.

Table 5.31 Matching genes in Ried et al rectal cancer data.

GENE NAME	ID
E2F2-E2F transcription factor 2 (E2F2)	H200012309
SLC25A5-solute carrier family 25	H200006643

Table 5.32 Comparison of accuracies obtained from different classifiers for predicting chemoradiotherapy response using the 3-gene signature. The improved overall accuracy of the prediction with the Decision stump classifier compared with other methods was assessed by significance testing (*N* = 17).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Decision stump	**90.00**	**57.10**	**73.55**	**76.47**	
Multiboost AB	90.00	28.60	59.30	64.70	<0.23
AdaboostM1	70.00	42.90	56.45	58.82	<0.14
Logisitc regression	70.00	28.60	49.30	52.94	<0.08
AD Tree	50.00	42.90	46.45	47.05	<0.04

Table 5.33 Confusion matrix obtained from the Decision stump classifier for predicting response using the 3-gene signature.

Actual/Predicted	a (response)	b (no response)
a (response)	9	1
b (no response)	3	4

5.3.3 Summary of validation results of 3-gene signature on rectal cancer data

Table 5.34 shows the details of validation results of 3-gene signature on rectal cancer data in different groups of files. For each dataset the classifier with the highest overall accuracy was reported. The validation results on cDNA data files, time-dependent ROC and Kaplan-Meier analyses were not reported as they were not significant.

Table 5.34 Summary of validation results of 3-gene signature on rectal cancer data.

Rectal cancer data	Classifier	Predicted variable	Sensitivity (%)	Specificity (%)	(Sensitivity + Specificity)/2 (%)	Overall accuracy (%)
Tumor biopsies 2 files	KStar	Lymph node status	83.33	60.00	71.66	76.47
Tumor biopsies 2 files	Decision stump	Chemoradiotherapy response	90.00	57.10	73.55	76.47

5.4 Validation results of the 5-gene signature on rectal cancer data

5.4.1 Predicting lymph node status in tumor biopsies 2 files

The tumor biopsies 2 data files from Ried et al (PMID 16397240) consisted of 17 patient samples. The search for matching genes was done using the gene symbols. There were 2 matching gene (Table 5.35). The data used for validation consisted of the expression of these 2 genes in the 17 patient samples. Weka software was used for validation and lymph node status (positive/negative) was predicted. Different classification schemes including Multiboost AB, Logitboost, Random Tree, Multilayer perceptron, and LWL were applied to this dataset to find the best scheme. Table 5.36 shows the comparison between Multiboost AB and some of the classifiers used for validation on other datasets. Multiboost AB classifier performed better than the other classifiers. It had a sensitivity of 83.30%, a specificity of 40.00%, and an overall accuracy of 70.58%. Table 5.37 shows the confusion matrix for Multiboost AB classifier. The difference in overall accuracy between Multiboost AB and other classifiers was not statistically significant due to the small sample size.

Table 5.35 Matching genes in Ried et al rectal cancer data.

GENE NAME	ID
TPD52L2-tumor protein D52-like2	H200013992
ZNF187-zinc finger protein 187	H200015602
SLC25A5-solute carrier family 25	H200006643

Table 5.36 Comparison of accuracies obtained from different classifiers for predicting lymph node status using the 5-gene signature. The improved overall accuracy of the prediction with the Multiboost AB classifier compared with other methods was assessed by significance testing (*N* = 17).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
Multiboost AB	**83.30**	**40.00**	**61.65**	**70.58**	
Adaboost M1	83.30	20.00	51.65	64.70	<0.36
Random Tree	58.30	20.00	39.15	47.05	<0.09
Random committee	75.00	20.00	47.50	58.82	<0.24
LWL	75.00	20.00	47.50	58.82	<0.24

Table 5.37 Confusion matrix obtained from the Multiboost AB classifier for predicting lymph node status using the 5-gene signature.

Actual/Predicted	**a (node positive)**	**b (node negative)**
a (node positive)	10	2
b (node negative)	3	2

5.4.2 Predicting chemoradiotherapy response in tumor biopsies 2 files

The tumor biopsies 2 data files from Ried et al (PMID 16397240) consisted of 17 patient samples. The search for matching genes was done using the gene symbols. There were 3 matching genes (Table 5.38). The data used for validation consisted of the expression of these 3 genes in the 17 patient samples. Weka software was used for validation and chemoradiotherapy response (yes/no) was predicted. Different classification schemes including *J48*, AD Tree, *IB1*, Logistic regression, and *AdaboostM1* were applied to this dataset to find the best scheme. Table 5.39 shows the comparison between *J48* and some of the classifiers used for validation on other datasets. *J48* classifier performed better than the other classifiers. It had a sensitivity of 90.00%, a specificity of 57.10%, and an overall accuracy of 76.47%. Table 5.40 shows the confusion matrix for *J48* classifier. The difference in overall accuracy between *J48* and Logistic regression (p < 0.04) was statistically significant.

Table 5.38 Matching genes in Ried et al rectal cancer data.

GENE NAME	ID
TPD52L2-tumor protein D52-like2	H200013992
ZNF187-zinc finger protein 187	H200015602
SLC25A5-solute carrier family 25	H200006643

Table 5.39 Comparison of accuracies obtained from different classifiers for predicting chemoradiotherapy response using the 5-gene signature. The improved overall accuracy of the prediction with the *J48* classifier compared with other methods was assessed by significance testing (*N* = 17).

Classifier	Sensitivity (%)	Specificity (%)	(Sensitivity+ Specificity)/2 (%)	Overall Accuracy (%)	*P*-value
J48	**90.00**	**57.10**	**73.55**	**76.47**	
AD Tree	80.00	42.90	61.45	64.70	<0.23
IB1	50.00	57.10	53.55	52.94	<0.08
Logistic regression	60.00	28.60	44.30	47.05	<0.04
AdaboostM1	70.00	42.90	56.45	58.82	<0.14

Table 5.40 Confusion matrix obtained from the *J48* classifier for predicting response using the 5-gene signature.

Actual/Predicted	**a (response)**	**b (no response)**
a (response)	9	1
b (no response)	3	4

5.4.3 Kaplan-Meier analyses on Ried et al rectal cancer data (*n=23*) using the 5-gene signature

The cDNA 2 files in rectal cancer data were checked for matching genes with the 5-gene signature. There were 3 matching genes. The Cox model based on the expression of these 3 genes was used to get recurrence risk scores for the 23 patients. The choices for choosing a cut-off value for patient stratification are the peak value from histogram, mean risk score or median risk score. In this analysis, the median risk score was chosen as cut-off as it resulted in best patient stratification. Cut-off values of 0.17 and -1.16 were chosen for relapse-free survival and overall survival, respectively. The *pamr* package in R was used to plot the Kaplan-Meier curves. The low-risk and high-risk groups, had distinct relapse-free survival (*p = 0.043, n=23,* log-rank tests), and distinct overall survival (*p = 0.036, n=23,* log-rank tests), respectively for the data in cDNA2 files. Table 5.41 shows the different parameters obtained from the Cox model using the 9-gene signature for disease-free survival and overall survival in cDNA 2 data files.

Table 5.41 Different parameters obtained from Cox model using the 5-gene signature for relapse-free survival and overall survival.

	Disease-free survival					Overall survival				
Gene Symbol	**coef**	**exp (coef)**	**se (coef)**	**z-score**	***p-value***	**coef**	**exp (coef)**	**se (coef)**	**z-score**	***p-value***
ZNF187	-0.2148	0.807	1.237	-0.1737	0.86	1.633	5.117	1.96	0.834	0.40
SLC25A5	-0.0006	0.999	0.825	-0.0007	1.00	-0.604	0.547	1.52	-0.398	0.69

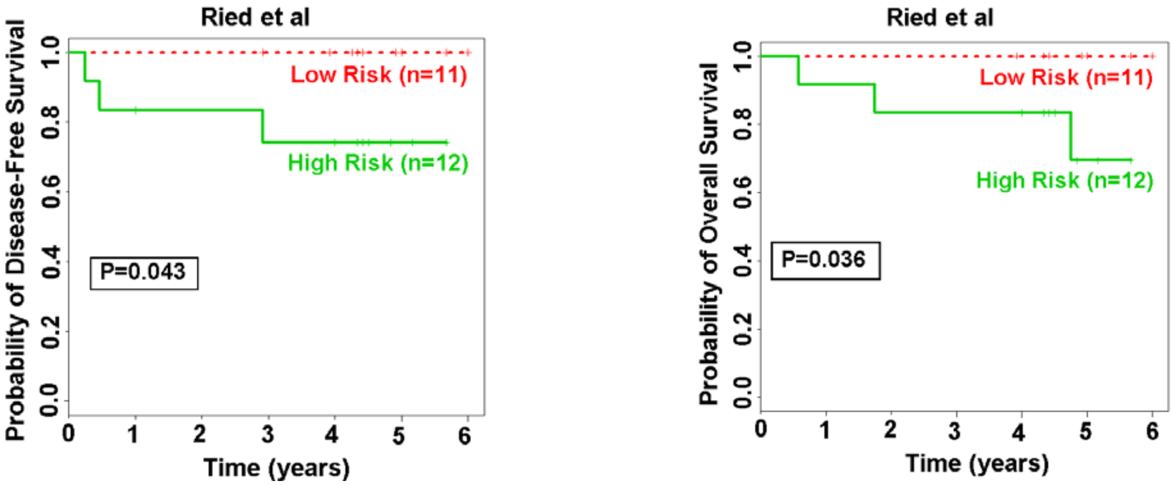

Figure 5.5 Kaplan-Meier plots on rectal cancer data (*n=23*) for disease-free survival and overall survival using the 5-gene signature in cDNA 2 data files.

5.4.4 Summary of validation results of 5-gene signature on rectal cancer data

Table 5.42 shows the details of validation results of 5-gene signature on rectal cancer data in different groups of files. For each dataset the classifier with the highest overall accuracy was reported. The time-dependent ROC, Kaplan-Meier analyses and validation results on cDNA data files were not reported as they were not significant.

Table 5.42 Summary of validation results of 5-gene signature on rectal cancer data.

Rectal cancer data	Classifier	Predicted variable	Sensitivity (%)	Specificity (%)	(Sensitivity + Specificity)/2 (%)	Overall accuracy (%)
Tumor biopsies 2 files	MultiboostAB	Lymph node status	83.30	40.00	61.65	70.58
Tumor biopsies 2 files	J48	Chemoradiotherapy response	90.00	57.10	73.55	76.47

5.5 Conclusion

The 9-gene lymph node status signature on the whole had optimal prediction accuracy on the rectal cancer data set. The 9-gene signature might be used for predicting lymph node status and chemoradiotherapy response in rectal cancer data and to stratify patients into low-risk and high-risk groups. The 3-gene signature can be used to predict lymph node status and chemoradiotherapy response of the patients in tumor biopsies 2 data. The 5-gene signature can be used to stratify patients into low-risk and high-risk groups, predict lymph node status and chemoradiotherapy response in tumor biopsies 2 data.

Chapter 6

6.1 Conclusions

The advents of high throughput technologies, such as DNA microarrays are revolutionalizing the field of medicine. DNA microarrays are a powerful means of monitoring thousands of gene expression levels at the same time. Machine learning techniques are playing a pivotal role in analyzing the generated microarray data. Recent studies have successfully applied the machine learning approaches to predict the cancer stage, treatment outcome, drug response, and promise treatments tailored to the patients. Presently there are no gene tests available for clinical usage in colon cancer while there are gene tests like MammaPrint and Oncotype DX for breast cancer prognosis. Our study was focused in the direction of identifying important biomarkers to predict colon cancer stage and recurrence, building prognostic models, and stratifying patients into low-risk and high-risk groups based on cDNA microarray data.

In an effort to overcome the limitations of the traditional staging systems, in the first part of our study a 9-gene lymph node status signature was identified by feature selection using random forests and then discarding genes without differential expression. A prognostic patient stratification scheme was developed based on this 9-gene signature using the Cox model. In the second part of the study, we focused on identifying biomarkers predicting recurrence. This was achieved by a combinatorial scheme employing feature selection using random forests in the first step and then using InfoGain feature selection method in the next step. Two recurrence gene signatures were identified and patient stratification schemes were developed based on these signatures to identify subgroups of patients, at low and high-risks of recurrence. Recurrence prediction models were built using classifiers in Weka software based on these gene signatures. The gene signatures identified in this study could be used for classifying new colon cancer patients into different stages of the disease and different prognostic risk groups.

The analysis of microarray gene expression data through machine learning methods currently faces two major problems. Firstly the high dimensionality of the feature space and secondly the fact that gene expression data are very noisy (26). Most of the machine learning algorithms have been developed for applications in domains, such as business, retail, and marketing. A typical data mining banking application, has thousands or millions of records, and at most a few hundred fields. In contrast, a microarray gene expression data may only have a few hundred records and thousands of fields. Also, majority of the techniques used in standard data mining applications are very sensitive to noise (1). We could solve the problem of high dimensionality to some extent by preprocessing the data and employing a combinatorial scheme for feature selection. In the future, there is a necessity for new machine learning techniques addressing the high dimensionality and noisy characteristics of microarray gene expression data. We faced with another problem of availability of colon cancer datasets. The number of colon cancer datasets publicly available is very less. They are not as widely available as lung cancer and breast cancer datasets. The colon cancer data used in our study was obtained from our research collaborator Dr.Ried. The survival information was not available for other colon cancer datasets used for validation and we could not perform the time-dependent ROC and Kaplan-Meier analyses. Availability of the survival information and more colon cancer datasets publicly in the future would allow for robust validation of the identified gene signatures providing us with more understanding of the results.

Reference List

1. Piatetsky-Shapiro G, Tamayo *P*: Microarray data mining: facing the Challenges. SIGKDD Explorations 2003, 5:1–5

2. Grade, M., et al. "Gene expression profiling reveals a massive, aneuploidy-dependent transcriptional deregulation and distinct differences between lymph node-negative and lymph node-positive colon carcinomas." Cancer Res. 67.1 (2007): 41-56.

3. Diaz-Uriarte, R. and Andres S. varez de. "Gene selection and classification of microarray data using random forest." BMC.Bioinformatics. 7 (2006): 3.

4. Adler, A. S., et al. "Genetic regulators of large-scale transcriptional signatures in cancer." Nat.Genet. 38.4 (2006): 421-30.

5. Bandres, E., et al. "A gene signature of 8 genes could identify the risk of recurrence and progression in Dukes' B colon cancer patients." Oncol.Rep. 17.5 (2007): 1089-94.

6. Graziano, F. and S. Cascinu. "Prognostic molecular markers for planning adjuvant chemotherapy trials in Dukes' B colorectal cancer patients: how much evidence is enough?" Ann.Oncol. 14.7 (2003): 1026-38.

7. Buyse, M. and *P*. Piedbois. "Should Dukes' B patients receive adjuvant therapy? A statistical perspective." Semin.Oncol. 28.1 Suppl 1 (2001): 20-24.

8. Wang, Y., et al. "Gene selection from microarray data for cancer classification--a machine learning approach." Comput.Biol.Chem. 29.1 (2005): 37-46.

9. Alon, U., et al. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays." Proc.Natl.Acad.Sci.U.S.A 96.12 (1999): 6745-50.

10. Ramaswamy, S., et al. "Multiclass cancer diagnosis using tumor gene expression signatures." Proc.Natl.Acad.Sci.U.S.A 98.26 (2001): 15149-54.

11. Tamayo, *P*., et al. "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." Proc.Natl.Acad.Sci.U.S.A 96.6 (1999): 2907-12.

12. Ian H. Witten, Eibe Frank. "Practical Machine Learning Tools and Techniques". second edition Morgan Kaufmann Publishers.

13. John, G. Cleary and Leonard, E. Trigg (1995) "K*: An Instance- based Learner Using an Entropic Distance Measure", Proceedings of the 12th International Conference on Machine learning, pp. 108-114.

14. G. Webb. "Multiboosting: A technique for combining boosting and wagging". Machine Learning, 40, 2000.

15. S. le Cessie and J. C. van Houwelingen."Ridge estimators in logistic regression". Appl Stat, 41:191–201, 1992.

16. Fox J. Cox proportional-hazard regression for survival data. "Appendix to An R and S-PLUS companion to applied regression". February 2002.

17. Bland, J. M. and D. G. Altman. "Survival probabilities (the Kaplan-Meier method)." BMJ 317.7172 (1998): 1572.

18. Jager, K. J., et al. "The analysis of survival data: the Kaplan-Meier method." Kidney Int. 74.5 (2008): 560-65.

19. Bland, J. M. and D. G. Altman. "The logrank test." BMJ 328.7447 (2004): 1073.

20. Heagerty, *P*. J., T. Lumley, and M. S. Pepe. "Time-dependent ROC curves for censored survival data and a diagnostic marker." Biometrics 56.2 (2000): 337-44.

21. Kwon, H. C., et al. "Gene expression profiling in lymph node-positive and lymph node-negative colorectal cancer." Dis.Colon Rectum 47.2 (2004): 141-52.

22. Koehler, A., et al. "Gene expression profiling of colorectal cancer and metastases divides tumours according to their clinicopathological stage." J.Pathol. 204.1 (2004): 65-74.

23. Croner, R. S., et al. "Microarray versus conventional prediction of lymph node metastasis in colorectal carcinoma." Cancer 104.2 (2005): 395-404.

24. Barrier, A., et al. "Colon cancer prognosis prediction by gene expression profiling." Oncogene 24.40 (2005): 6155-64.

25. Barrier, A., et al. "Prognosis of stage II colon cancer by non-neoplastic mucosa gene expression profiling." Oncogene 26.18 (2007): 2642-48.

26. Daniel *P*, Brian S, Ian B, Werner D. "Microarray data integration and machine learning techniques for lung cancer survival prediction." School of biomedical sciences, University of Ulster at Coleraine, Northern Ireland.

27. Koinuma, K., et al. "Epigenetic silencing of AXIN2 in colorectal carcinoma with microsatellite instability." Oncogene 25.1 (2006): 139-46.

28. Barrier, A., et al. "Stage II colon cancer prognosis prediction by tumor gene expression profiling." J.Clin.Oncol. 24.29 (2006): 4685-91.

29. Perou, C. M., et al. "Molecular portraits of human breast tumours." Nature 406.6797 (2000): 747-52.

30. Sorlie, T., et al. "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications." Proc.Natl.Acad.Sci.U.S.A 98.19 (2001): 10869-74.

31. van, 't, V, et al. "Gene expression profiling predicts clinical outcome of breast cancer." Nature 415.6871 (2002): 530-36.

32. Grade, M., et al. "Aneuploidy-dependent massive deregulation of the cellular transcriptome and apparent divergence of the Wnt/beta-catenin signaling pathway in human rectal carcinomas." Cancer Res. 66.1 (2006): 267-82.

33. Wang, Y., et al. "Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer." J.Clin.Oncol. 22.9 (2004): 1564-71.

34. Ma, Y., et al. "Population-based molecular prognosis of breast cancer by transcriptional profiling." Clin.Cancer Res. 13.7 (2007): 2014-22.

35. Bhatavdekar, J. M., et al. "Molecular markers are predictors of recurrence and survival in patients with Dukes B and Dukes C colorectal adenocarcinoma." Dis.Colon Rectum 44.4 (2001): 523-33.

36. Frank, E., et al. "Data mining in bioinformatics using Weka." Bioinformatics. 20.15 (2004): 2479-81. 4.

37. Artinyan, A., et al. "Molecular predictors of lymph node metastasis in colon cancer: increased risk with decreased thymidylate synthase expression." J.Gastrointest.Surg. 9.9 (2005): 1216-21.

38. E. Bair, R. Tibshirani, “Machine learning methods applied to DNA microarray data can improve the diagnosis of cancer,” SIGKDD Explorations, vol. 5(2), pp. 48-55, 2003.

39. Aliferis, C. F., D. Hardin, and P. P. Massion. "Machine learning models for lung cancer classification using array comparative genomic hybridization." Proc.AMIA.Symp. (2002): 7-11.

40. Strobl, C., et al. "Bias in random forest variable importance measures: illustrations, sources and a solution." BMC.Bioinformatics. 8 (2007): 25.

41. Ma, Y., et al. "Predicting cancer drug response by proteomic profiling." Clin.Cancer Res. 12.15 (2006): 4583-89.