Finding Biomarkers for Non-Small Cell Lung Cancer Diagnosis with Novel Data Mining Techniques

Quoc-Nam Tran†, Lamar (Texas State) University, USA.

***Abstract*—Non-small cell lung carcinoma (NSCLC) is the most common cause of worldwide cancer premature death with a very low survival rate of 8%-15%. Patients with an early stage diagnosis can have up to four times the survival rate of 40%-55%. Hence, discovering cost-effective biological markers that can be used to improve the diagnosis and prognosis of the disease is an important clinical challenge.**

Significant progress has been made to address this challenge. Some sets of biomarkers were identified in the last few years ranging from 5-gene signatures to 133-gene signatures. Since datasets of gene-expression profiles typically have tens of thousands of genes for just few hundreds of patients, this type of datasets will create many technical challenges impacting the accuracy of the diagnostic prediction. A typical molecular sub-classification method for lung carcinomas would have a low predictive accuracy of 68%-71%.

In this paper, we present a new data mining method that finds genetic markers and uses the markers to predict with up to 100% accuracy whether a patient has NSCLC and the sub-type of cancer in case the patient has NSCLC. Our method overcomes many challenges arose from datasets of gene-expression profiles. The new method discovers novel genetic changes that occur in lung tumors using gene-expression profiles. We discovered that a small set of nine gene-signatures (JAG1, MET, CDH5, ABCC3, DSP, ABCD3, PECAM1, MAPRE2 and PDF5) from the dataset of 12,600 gene-expression profiles of NSCLC acts like an inference basis for NSCLC lung carcinoma and hence can be used as genetic markers. This very small and *previously unknown set of biological markers* gives an almost perfect predictive accuracy for the diagnosis of the disease.

While proteins encoded by some of these gene-signatures (e.g., JAG1 and MAPRE2) have been showed to involve in the signal transduction of cells and proliferative control of normal cells, specific functions of proteins encoded by other gene-signatures have not yet been determined. Therefore, this work opens new questions for structural and molecular biologists about the role of these gene-signatures for the disease.

† Supported in part by NSF award CCF-0917257.

***Keywords*-Mining gene-expression profiles in bioinformatics, lung cancer, diagnosis.**

I. Introduction

In the last several years, one in four deaths in the United States is due to cancer, which makes cancer a major public health problem in the United States as well as many other parts of the world [1, 2]. Currently, cancer is a leading cause of death in the United States, second only to cardiovascular diseases. Last year, 1.48 million people were diagnosed with cancer, and 562,340 people died from cancer. The top five most common cancer-related deaths were due to lung, breast, prostate, colorectal and pancreatic cancer. Together, these five diseases accounted for over 50% of all cancer deaths in the United States in 2009. Lung cancer alone, with NSCLC as the most common cause of worldwide cancer premature death, killed over 160,000 people, more than the other four cancers put together. The disease has a very low survival rate of 8%-15%. Meanwhile, the survival rate for patients with early-stage disease increases to 40%-55% after surgery. That said, discovering cost-effective biological markers that can be used to improve the diagnosis and prognosis of the disease is an important clinical challenge [3].

NSCLC is sub-categorized as adenocarcinomas, squamous cell carcinomas, and large-cell carcinomas, of which adenocarcinomas are the most common [4]. The histopathological sub-classification of lung adenocarcinoma is challenging. For example, in one study independent lung pathologists agreed on lung adenocarcinoma sub-classification in only 41% of cases [5]. In another study, proportional hazard models identified an optimal set of 50 prognostic mRNA transcripts using a 5-fold cross-validation procedure. This signature was tested in an independent set of 36 squamous cell lung carcinomas (SCC) samples and achieved 84% specificity and 41% sensitivity with an overall predictive accuracy of 68%

[6]. Combining the SCC classifier with their adenocarcinoma prognostic signature gave a predictive accuracy of 71% in 72 NSCLC samples.

Multiple techniques have evolved over the past few years allow rapid measurement of gene expression and simultaneous high-throughput measurement of thousands of genes from several hundred samples. Different parts of the gene-protein relationship can be measured such as messenger RNA levels, protein expression and cellular metabolic activity. Some of the available genomic technologies include gene expression arrays, serial analysis of gene expression, single-nucleotide polymorphism analysis, and high-throughput capillary sequencing [3].

Gene-expression array analysis methodologies developed over the last few years have demonstrated that expression data can be used in a variety of class discovery or class prediction biomedical problems including those relevant to tumor classification [7, 8, 9, 10]. Data mining and statistical techniques applied to gene expression data have been used to address the questions of distinguishing tumor morphology, predicting post treatment outcome, and finding molecular markers for disease [11, 12, 13, 14].

However, gene expression profiles present many challenges for data mining both in finding differentially expressed genes, and in building predictive models because the datasets are highly multidimensional (12,600 dimensions in our study) and contain a small number of records (197 records in our study). Although microarray analysis tool can be used as an initial step to extract most relevant features, one has to avoid overfitting the data and deal with the very large number of dimensions of the datasets. The current challenges in analyzing gene-expression profiles, is illustrated in a method recently published in the Journal of Experimental & Clinical Cancer Research in July 2009 [15] where it used prior knowledge with support vector machine-based classification in diagnosis of lung cancer. The authors of [15] reported an accuracy of 98.51%-99.06% for their classification algorithm using 5 marker genes on a dataset of 31 malignant pleural mesothelioma (MPM) and 150 lung adenocarcinomas. Even though the method in [15] can differentiate between MPM and lung adenocarcinomas with high accuracy, it gives an accuracy of 70% when we added other types of NSCLC lung cancer including adenocarcinomas, squamous cell lung carcinomas and pulmonary carcinoids into consideration. Other researchers also limited themselves in differentiate two sub-types of NSCLC lung cancer such as between adenocarcinomas and squamous cell lung carcinomas.

This paper aims at a novel data mining method that finds cost-effective genetic markers and uses the markers to differentiate with very high accuracy *all sub-types of NSCLC lung cancer*. Comparing with a recent publication [16] in that the author uses currently available data mining techniques in Weka to find biomarkers for NSCLC lung cancer, we found that our new method finds significantly more cost-effective genetic markers and provides more accurate sub-classification of NSCLC lung cancer. Comparison with SAM [17], a popular method for significance analysis of microarrays, is also provided in Section III.

Among the nine gene-signatures found by our new method (JAG1, MET, CDH5, ABCC3, DSP, ABCD3, PECAM1, MAPRE2 and PDF5), proteins encoded by some of these gene-signatures (e.g., JAG1 and MAPRE2) have been showed to involve in the signal transduction of cells and proliferative control of normal cells [18]. It has also been found that MAPRE2 is highly expressed in pancreatic cancer cells, and seems to be involved in perineural invasion [19]. However, specific functions of proteins encoded by other gene-signatures have not yet been determined. Hence, this work opens new questions for structural and molecular biologists about the role of these gene-signatures for the disease.

II. A New Data Mining Method for Significant Genes Selection & Sub-Classification

Before presenting our new algorithm for finding genetic markers and predicting NSCLC lung cancer, we will address the challenges one has to overcome while working with gene-expression profile datasets. Basic information about Gini indexes and classification algorithms can be found in many data mining books [20, 21, 22].

A. Solving the bias due to the order of classes

The first challenge that arose from the gene-expression datasets is the bias due to the order of cancer types or classes in data mining's terminology. Let's consider a simple example of expression profiles for a gene A in Table I where the gene dataset D has $d = 100$ elements

Range/Class	C_1	C_2	C_3
R_1	4	6	30
R_2	6	30	4
R_3	0	4	16

Table I
BIAS DUE TO THE ORDER OF CLASSES

and three classes. The gene expression values were partitioned into three ranges. Clearly, the cancer types or classes can be labeled in any order. When this gene is ranked by current microarray analysis methodologies, for example by calculating the Gini index $gini_A(D) = \sum_{i=1}^{m} \frac{|R_i|}{d} \cdot gini(R_i)$, the first two rows contribute equally to the Gini index because $gini(R_i) = 1 - \sum_{j=1}^{n} p_{i,j}^2$ where $p_{i,j} = \frac{|C_{i,j}|}{|R_i|}$ is the relative frequency of class C_j in R_i, and $|\cdot|$ is the notation for cardinality [23]. We have the same problem when entropy is calculated instead of the Gini index. That said, when one just considers the probability distribution without taking into account the order of the classes, the first two partitions of expression profiles will contribute equally. Clearly, the two partitions should contribute differently because Partition R_1 says that 75% of patients with gene expression values within this range are classified into Class C_3 while Partition R_2 says that 75% of patients with gene expression values within this range are classified into Class C_2. Hence, in order to have a robust gene selection method, one has to differentiate the partitions with different class orders because they have different amount of information.

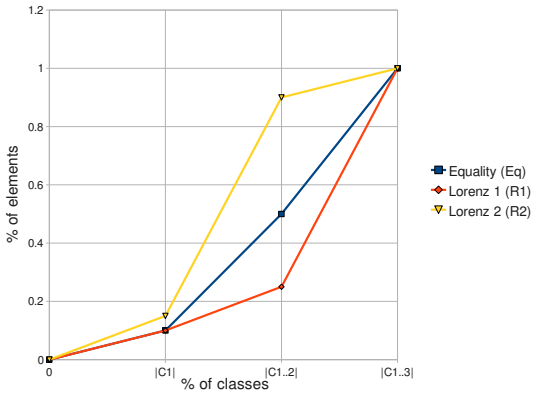

Figure 1. Lorenz curves

To solve this problem, we generalized the well known Lorenz curves, a common measure in economics to gauge the inequalities in income and wealth. In Figure 1, we illustrate how modified Lorenz curves and modified Gini coefficients are calculated. The Equality Polygon (Eq) is defined based on the percentages of elements in $|C_1|$, $|C_{1..2}| = |C_1| + |C_2|$, ..., $|C_{1..n}| = \sum_{j=1}^{n} |C_j|$ at $x-$coordinates 0, $1/n$, $2/n$, ..., 1, where n is the number of classes and $|C_1| \leq |C_2| \leq$, ..., $\leq |C_n|$. The Lorenz polygon of a partition, say R_i, is defined based on the percentage of elements in $|C_{i,1}|$, $|C_{i,1}| + |C_{i,2}|$, ..., $\sum_{j=1}^{n} |C_{i,j}|$ at $x-$coordinates 0, $1/n$, $2/n$, ..., 1. The Gini coefficient of a partition, say R_i, is defined as $(\int_0^1 L(R_i) \cdot dx - \int_0^1 Eq \cdot dx) / \int_0^1 Eq \cdot dx$. One can easily see that the partitions with different class orders are now differentiated.

B. Solving the bias due to the order of gene expression values

Another technical challenge for microarray analysis methodologies comes from the order of discretized gene expression values. Let's consider another simple example

Class/ Range	C_1	C_2	C_3
R_1	3	0	0
R_2	0	100	0
R_3	4	0	0
R_4	0	0	5

Class/ Range	C_1	C_2	C_3
R_1	3	0	0
R_2	4	0	0
R_3	0	100	0
R_4	0	0	5

Table II
BIAS DUE TO THE ORDER OF GENE EXPRESSION VALUES

of gene-expression profiles for two genes in Table II with three classes. The gene expression values were discretized into four ranges. In contrast to the previous challenge, the ranges of gene-expression values do follow some order. When this genes are ranked by current microarray analysis methodologies, for example by calculating the Gini index of gene A using dataset D $gini_A(D) = \sum_{i=1}^{m} \frac{|R_i|}{d} \cdot gini(R_i)$ where $d = |D|$, the two genes would have the same rank. Clearly, the gene-expression profiles on the right hand side of Table II have a more harmonic distribution with respect to the rows in comparison with the gene on the left. That said, these two genes should be ranked differently.

To solve this problem, we generalized the Gini coefficients by taking into account the splitting status and the Gini ratio. The splitting status of D with respect to the attribute A is calculated as

$$split_A(D) = 1 - \sum_{i=1}^{m} (\frac{|R_i|}{d})^2.$$

The Gini ratio of D with respect to the attribute A is defined as $LorenzGini(A) = \Delta gini(A) / split_A(D)$, where $\Delta gini(A) = gini(D) - gini_A(D)$ and $gini(D) = 1 - \sum_{j=1}^{n} (\frac{|C_j|}{d})^2$.

Furthermore, to take into account the gene expression profiles with different value orders, the Gini coefficient is calculated as $gini_A(D) = \sum_{i=1}^{m} \frac{|R_i|}{d} \cdot \delta(i) \cdot gini(R_i)$, where $\delta(i)$ is the sum of the normalized distances between the row i and rows $i-1$, $i+1$. The coefficient $\delta(i)$ is used as a weight to emphasize a row when it is close to its neighbors.

C. New Algorithm

Input: A gene-expression profiles dataset D with up to 34,000 dimensions.
Output: A small subset of genes as genetic markers and a prediction model for NSCLC lung cancer
Step1: Discretize the gene-expression profile values.
Step2: Select genetic markers by using the genes with highest ranking LorenzGini.
Step3: Build the prediction model to classify patients using the genetic markers.

A threshold can be used for controlling the number of significant genes for genetic markers. The splitting status of dataset D with respect to a gene A can be calculated as a by-product when the reduction in impurity of D with respect to the attribute A is calculated. Therefore, the time complexity and space complexity of the algorithm are the same as the complexities of Gini index algorithm.

Our method has been implemented in Maple and Weka [24, 25]. In the next section, we will present our experiment with a dataset of gene-expression profiles of NSCLC from the mRNA expression profiles.

Notice that our new method works for any dataset with $\geq$ 2 classes. For any number of classes, even when the number of classes is equal to 2, the new method is completely different with other microarray analysis methodologies.

III. Experimentation

A. mRNA Materials

To test and validate our algorithm, we extract the gene-expression profiles of NSCLC from the mRNA expression profiles in [26] in that a total of 203 snap-frozen lung tumors (n=186) and normal lung (n=17) specimens were used to create the dataset. Of these, 125 adenocarcinoma samples were associated with clinical data and with histological slides from adjacent sections. The 203 specimens include histologically defined lung adenocarcinomas (n=139), squamous cell lung carcinomas (n=21), pulmonary carcinoids (n=20), and normal lung (n=17) specimens. Total RNA extracted from samples was used to generate cRNA target, subsequently hybridized to human U95A oligonucleotide probe arrays according to standard protocols. As the result, we obtained a dataset of 12,600 gene-expression profiles for 197 patients.

B. Finding genetic markers

Using the algorithm described in the previous section, we select 250 genes with the highest LorenzGini indexes. To further reduce the size of the gene subsets and to improve the prediction accuracy, we evaluate different combinations of genes to identify an optimal subset in terms of accuracy for the Bayesian Net classification. The gene subsets to be evaluated are generated using different subset search techniques. We use Best First and Greedy search methods in the forward and backward directions. Greedy search considers changes local to the current subset through the addition or removal of genes. For a given parent set, a greedy search examines all possible child subsets through either the addition or removal of genes. The child subset that shows the highest goodness measure then replaces the parent subset, and the process is repeated. The process terminates when no more improvement can be made. Best First search is similar to greedy search in that it creates new subsets based on the addition or removal of genes to the current subset with the ability to backtrack along the subset selection path to explore different possibilities when the current path no longer shows improvement. To prevent the search from backtracking through all possibilities in the gene space, a limit is placed on the number of non-improving subsets that are considered. In our evaluation we chose a limit of five.

The algorithm returns a set of nine genes (JAG1, MET, CDH5, ABCC3, DSP, ABCD3, PECAM1, MAPRE2 and PDF5) from the dataset of 12,600 gene-expression profiles of NSCLC. We exploit this small set of genes to differentiate all sub-types of NSCLC lung cancer.

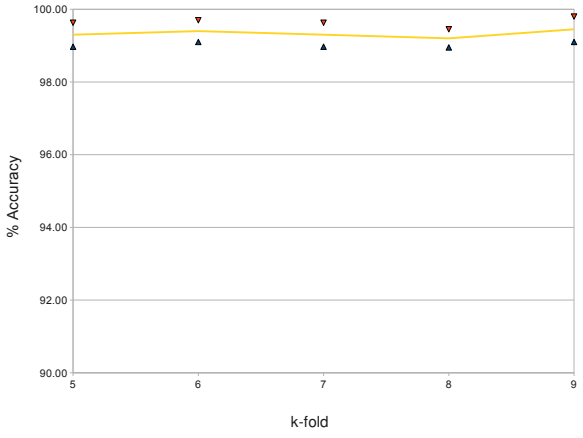

Figure 2. Accuracy of sub-classifications with standard deviations

To build the classification model, we used Bayesian Network (BayesNet), which is structured as a combination

of a directed acyclic graph of nodes and links, and a set of conditional probability tables. Nodes represent features or classes, while links between nodes represent the relationship between them. Conditional probability tables determine the strength of the links. There is one probability table for each node (feature) that defines the probability distribution for the node given its parent nodes. If a node has no parents the probability distribution is unconditional. If a node has one or more parents the probability distribution is a conditional distribution, where the probability of each feature value depends on the values of the parents.

Figure 2 shows the averaged accuracies of the gene expression profile classification using Bayesian Net classification together with their standard deviations. To test the accuracy of classification models, we use k-fold cross validation, which is a common method for estimating the error of a model on benchmark medical data sets. The reason for using this testing approach is that when a model is built from training data, the error on the training data is a rather optimistic estimate of the error rates the model will achieve on unseen data. The aim of building a model is usually to apply the model to new, unseen data– we expect the model to generalize to data other than the training data on which it was built. Another reason for using this testing approach is that the available medical data sets are small and no test data set is available. It is well-known that k-fold cross-validation is very useful for this type of data sets.

For a reliable evaluation of the accuracy, we test the classification algorithm for many values of k. More precisely, we test for $k = 5..9$. For each value of k, the data set D is randomly divided into k subsets D_1, D_2 , ..., D_k. We leave out one of the subsets D_i, $i = 1..k$ each time for being used as a test data set for cross validation. The remaining subset $\cup_{j \neq i} D_j$ is used to build the model. The cross validation accuracy computed for each of the k test samples are then accumulated to give the k-fold estimate of the cross validation accuracy. To ease the effects of the random partitions on the data set, this whole process is repeated 10 times with different random seeds and the results are then averaged to give the estimated accuracy of the comparing algorithms in Figure 2.

During the validation process, all patients with lung adenocarcinomas were correctly predicted, all patients except one with squamous cell lung carcinomas were correctly predicted, all patients with pulmonary carcinoids were correctly predicted, and all patients with normal lung specimens were correctly predicted. The only false prediction for random seed 1 was a patient with squamous cell lung carcinomas but incorrectly predicted as adenocarcinomas. As we can see, this very small set of genes gives an almost perfect predictive accuracy for the diagnosis of the disease. When the number of genes is further reduced or increased, the accuracy starts to declined. That said, this set of nine genes acts like an inference basis for NSCLC lung carcinoma and hence can be used as genetic markers.

C. Comparing with other gene selection methods

We now investigate the classifying accuracy of the significant genes generated by LorenzGini with respect to the size of the reduced microarray datasets. Comparing with a recent publication [16] in that the author uses currently available data mining techniques in Weka to find biomarkers for NSCLC lung cancer, we found that our new method finds significantly more cost-effective genetic markers and provides more accurate sub-classification of NSCLC lung cancer. We also compare our method with SAM using the same dataset for NSCLC lung cancer. SAM combines t-test and permutations to calculate a False Discovery Rate to provide a subset of genes that are considered significant [17]. Using SAM, we select four sets of 50, 100, 150, 200 and 250 most significant genes by using the parameter values of 0.556. 0.458. 0.4188, 0.383 and 0.3568, respectively.

We then use the Bayesian Net classification in Weka to check the accuracy of the most significant gene sets generated by LorenzGini and SAM [25]. Besides our fresh implementation of LorenzGini algorithms, simple converters were written to connect SAM and Weka. For a reliable evaluation of the accuracy, we test the classification algorithm for many values of k as specified in our validation plan.

Figure 3 shows the accuracy of the gene expression profile classification using Bayesian Net algorithm on SAM's gene sets and on LorenzGini's gene sets with 50 genes. As we can see, the classifying accuracy has been improved with the LorenzGini's gene selections. We also observed that the accuracy of the gene expression profile classification using Bayesian Net algorithm on SAM's gene sets declined when the number of genes is reduced to 50 and below. In contrast, the accuracy of the gene expression profile classification using LorenzGini's gene sets is stable even when the number of genes is reduced to 9, which has the highest accuracy. This observation is also true for other classification methods.

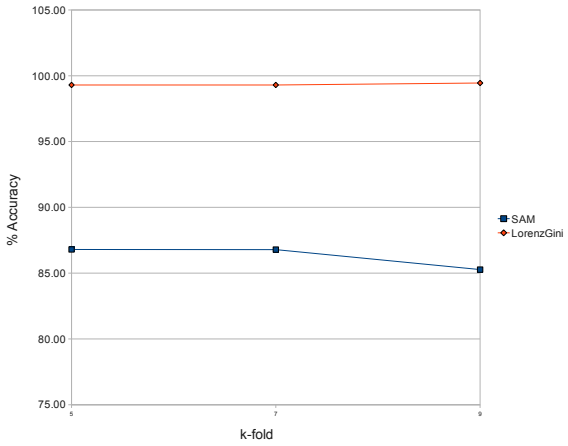

Figure 3. SAM's & LorenzGini's gene sets classified by Bayesian Net

IV. Conclusion

We presented a method that can find cost-effective biological markers as quantifiable measurements for an almost perfect predictive accuracy of NSCLC lung cancers. As cancers are complicated, one can only predict the status using a combination of many genes. The genes we discovered as genetic markers (JAG1, MET, CDH5, ABCC3, DSP, ABCD3, PECAM1, MAPRE2 and PDF5) are different with previously known results. Furthermore, proteins encoded by some of these gene-signatures (e.g., JAG1 and MAPRE2) have been showed to involve in the signal transduction of cells and proliferative control of normal cells while specific functions of proteins encoded by other gene-signatures have not yet been determined. Therefore, this work opens new questions for structural and molecular biologists about the role of these gene-signatures for the disease.

References

[1] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, and M. J. Thun, "Cancer statistics, 2007," *CA Cancer J Clin*, vol. 57, pp. 43–66, 2007.

[2] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun, "Cancer statistics, 2009," *CA Cancer J Clin*, vol. 59, pp. 225–249, 2009.

[3] S. Singhal, D. Miller, S. Ramalingam, and S.-Y. Sun, "Gene expression profiling of non-small cell lung cancer," *Lung cancer*, vol. 60, no. 3, pp. 313–324, 2008.

[4] J. D. Watson, "The human genome project: past, present, and future," *Science*, vol. 248, pp. 44–49, 1990.

[5] F. S. Collins, M. Morgan, and A. Patrinos, "The human genome project: lessons from large-scale biology," *Science*, vol. 300, pp. 286–290, 2003.

[6] B. Cox, T. Kislinger, and E. A., "Integrating gene and protein expression data: pattern analysis and profile mining," *Methods*, vol. 35, no. 3, pp. 303–314, 2005.

[7] A. Butte, "The use and analysis of microarray data," *Nature Review Drug Discovery*, vol. 1, no. 12, pp. 951–960, 2002.

[8] G. Piatetsky-Shapiro and P. Tamayo, "Microarray data mining: Facing the challenges," *SIGKDD Explorations*, vol. 5, no. 2, 2003.

[9] S. Ramaswamy and T. R. Golub, "DNA microarrays in clinical oncology," *Journal of Clinical Oncology*, vol. 20, pp. 1932–1941, 2002.

[10] P. Tamayo and S. Ramaswamy, "Cancer genomics and molecular pattern recognition," in *Expression profiling of human tumors: diagnostic and research applications*, M. Ladanyi and W. Gerald, Eds. Humana Press, 2003.

[11] W. Dalton and S. Friend, "Cancer biomarkers–an invitation to the table," *Science*, vol. 312, no. 5777, pp. 1165–1168, 2006.

[12] T. J. Yeatman, "Predictive biomarkers: Identification and verification," *J Clin Oncol*, vol. 27, no. 17, pp. 2743–2744, 2009.

[13] K. Shedden, J. Taylor, S. Enkemann, M. Tsao, T. Yeatman, W. Gerald, S. Eschrich, I. Jurisica, T. Giordano, D. Misek, A. Chang, C. Zhu, S. D., S. Hanash, F. Shepherd, K. Ding, L. Seymour, K. Naoki, N. Pennell, B. Weir, R. Verhaak, C. Ladd-Acosta, T. Golub, M. Gruidl, A. Sharma, J. Szoke, M. Zakowski, V. Rusch, M. Kris, A. Viale, N. Motoi, W. Travis, B. Conley, V. Seshan, M. Meyerson, R. Kuick, K. Dobbin, T. Lively, J. Jacobson, and D. Beer, "Gene expression-based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study," *Nat Med*, vol. 14, pp. 822–827, 2008.

[14] B. Kim, H. J. Lee, H. Y. Choi, Y. Shin, S. Nam, G. Seo, D.-S. Son, J. Jo, J. Kim, J. Lee, J. Kim, K. Kim, and S. Lee, "Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data," *Cancer Res*, vol. 67, pp. 7431–8, 2007.

[15] P. Guan, D. Huang, M. He, and B. Zhou, "Lung cancer gene expression database analysis incorporating prior knowledge with support vector machine-based classification method," *J Exp Clin Cancer Res.*, vol. 28, no. 103, pp. 1–7, 2009.

[16] N.-P. Tran, "Using data mining techniques for improving non-small cell lung cancer classification," *Journal of Computing Sciences in Colleges*, 2010, accepted for publication.

[17] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Natl. Acad. Sci. USA*, vol. 98, pp. 5116–5121, 2001.

[18] F. J., A. Chung, H. Xu, J. Zhu, H. Outtz, J. Kitajewski, Y. Li, X. Hu, and L. Ivashkiv, "Autoamplification of notch signaling in macrophages by tlr-induced and rbp-j-dependent induction of jagged1," *J Immunol*, vol. 185, no. 9, pp. 5023–31, 11 2010.

[19] A. I., G. S., D. T., K. T., B. K., G. N.A., F. H., and K. J., "The microtubule-associated protein mapre2 is involved in perineural invasion of pancreatic cancer cells," *Int J Oncol*, vol. 35, no. 5, pp. 1111–6, 2009.

[20] J. Han and K. Micheline, *Data Mining: Concepts and Techniques*, 2nd ed. Morgan Kaufmann, 2006.

[21] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, 2008.

[22] N. Ye, Ed., *The Handbook of Data Mining*. Lawrence Erlbaum Associates, 2003.

[23] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software, 1984, monterey, CA.

[24] B. W. Char, K. O. Geddes, G. H. Gonnet, B. L. Leong, M. B. Monagan, and S. M. Watt, *Maple V Language Reference Manual*. Springer Verlag, 1991.

[25] "http://www.cs.waikato.ac.nz/ml/weka," 2009.

[26] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, T. R. Golub, D. J. Sugarbaker, and M. Meyerson, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 24, pp. 13 790–13 795, 2001.