Data Mining on Survival Prediction after Chemotherapy for Diffuse Large-B-Cell Lymphoma and Genomics of Metastasis Cancer

Shen LU  
Soft Challenge LLC  
Little Rock, AR, USA  
slu@softchallenge.net

Richard SEGALL  
Arkansas State University at Jonesboro  
Jonesboro, AR, USA  
rsegall@astate.edu

Thomas HAHN  
University of Arkansas at Little Rock  
Little Rock, AR, USA  
tfahn@ualr.edu

ABSTRACT
This research pertains to the applications of data mining of microarray databases for large-B-cell Lymphoma and metastasis cancer, the latter of which little has been known about the genomic events that regulate the transformation of a tumor into a metastatic phenotype.

1. INTRODUCTION
Microarray technology has found its applications in recent years in many fields of life science. Generally speaking, all the data analysis behind these applications can be characterized into two major categories: (i.) discovery and (ii.) prediction. Discovery is to discover new knowledge, new genes involved in a pathway; prediction is to create predictive models to be used in such areas as toxicology and disease diagnosis. Fundamental to both discovery and prediction is the selection of genes that are differentially expressed (up or down) when comparing the samples of your interest to the control group.

Both discovery and prediction can help make diagnosis in the perspective of the lab research. Microarray analysis should be consistent with the clinical diagnosis. If both of them have the same conclusion, the diagnostic explanation can be accurate with a high probability; but on the other hand, if their conclusions conflict with each other, neither of them can be useful. In this paper, we use data mining techniques to build prediction models using microarray expression data. After that, we further check with the clinical gene signatures in order to find out if the significant genes that can be used to make prediction models for a particular disease, such as lymphoma, are in gene signature which is built based on clinical predictors, such as international prognostic index (IPI).

2. RELATED WORK
The authors Lu and Segall have performed many previous studies on applications of data mining to microarray databases as evidence by references to related work outside the scope of this paper. Lu and Segall [(2011), (14), (2011), (16)] for application of statistical quality control of microarray gene expression, Lu et al. [(2013), (16), (2013), (17)] for comparison of data mining methods on microarray gene expression data on cancer, and Lu et al. (2013), (18)] as a poster of preliminary research of this paper. Segall (2006), (23) (was one of the first publications in the area of data mining of microarray databases for biotechnology. Segall [(2005), (24), (2005), (25)] performed data mining of environmental factors on plants. Segall and Pierce [(2009)], (26), (2009), (27)] discussed data mining of leukemia cell microarrays and Segall and Pierce [(2009)], (28) extended these using self-organized maps. Segall and Zhang [(2007)], (29), (2006), (30), (2008), (31)] performed data mining for human lung cancer and other.

Wright et al. (2003), (34)] used Bayes’ rule to classify diffuse large B cell lymphoma (DLBCL) biopsy samples into two gene expression subgroups based on data obtained from spotted cDNA microarrays. They next used this predictor to discover these subgroups within a second set of DLBCL biopsies that had been profiled by using oligonucleotide microarrays. They identified the germinal center B-like (GCB) and activated B-cell like (ABC) DLBCL subgroups which have significantly different 5-yr survival rates after multiagent chemotherapy (62% vs. 26%; p=0.0051), in accordance with the analysis of other DLBCL cohorts.

Wright and Simon (2003), (35) proposed a model which can be used to draw gene variances from an inverse gamma distribution and estimate parameters afterwards. The motivation of their work is that DLBCL dataset has limited samples which makes estimation difficult since variance estimates made on a gene by gene basis will have few degree of freedom and the assumptions that all genes share equal variance is unlikely to be
true. This model results in a test statistic that is a minor variation of those used in standard linear models and has more power than standard tests to pick up large changes in expression and does not increase the rate of false positives.

Ein-Dor et al. ((2005), [9]) performed research into the overlap genes of microarray expression data in order to find out whether the different results of the same genes are because of different technologies, or because of different patients and different types of analyses. They used a simple method to experiment on a breast cancer microarray dataset. The result set of the genes are not unique which is strongly influenced by the subset of patients used for gene selection.

Colomo et al. ((2003), [6]) concluded that microarray gene expression profiling is associated with particular clinicopathological features but is not essential to predicting outcome in DLBCL patients.

Ross et al. ((2003), [22]) demonstrated that expression profiling of leukemic blasts can accurately identify all of the known prognostic subtypes. By analyzing the leukemic blasts microarray gene samples, the newly identified subtype discriminating genes are novel markers for those not identified in previous study. The newly selected genes are highly ranked as class discriminators that have not yet been used and should be used in clinical trials.

Hans et al. ((2004), [11]) divided diffuse large-B-cell lymphoma into prognostically important subgroups with germinal center B-cell like, activated B cell like and type 3 gene expression profiles using a cDNA microarray of the created tissue microarray blocks. They concluded that immunostains can be used to determine the GCB and non-GCB subtypes of DLBCL and predict survival similar to the cDNA microarray.

3. BACKGROUND

3.1 Microarray Profiling

For two-color microarray experiments, as shown in Figure 1, one must decide what the most appropriate comparison is to be made for each array of hybridization. The simplest comparisons can be separated into four general classes, such as direct comparison, reference design, balanced block design and loop design. In many ways, direct comparisons are the simplest conceptually; they are used when two distinct classes of experimental samples are to be compared, such as a treated sample and its untreated control. On each array, representatives of the two classes are paired and co-hybridized together such that the relative expression levels are measured directly on each array. The choice of appropriate pairing depends on the experimental question under study. For example, one can pair diseased and normal tissue from the same patient or randomly select animals from mutual and wild-type groups.

The strategy to collect data for any given case is influenced by a wide range of factors, including the availability of samples, the quantity of RNA that can be obtained, the size of the study, and the logistical constraints in the laboratory.

For each gene, the process begins with defining an expression vector that represents its location in expression space. In this view of gene expression, each hybridization represents a separate distinct axis in space, and the log2(ratio) measured for that gene in that particular hybridization represents its geometric coordinate. In this way, expression data can be represented in m-dimensional expression space, where m is the number of hybridizations and where each gene expression vector is represented as a single point in that space. It should be noted that one could use a similar approach to representing each hybridization assay using a sample vector consisting of the expression values for each gene; these define a sample space whose dimension is equal to the number of genes assayed in each array.

Figure 1: Illustration of a microarray containing thousands of “spots” of genomic data [2]

3.2 Data Mining using Self-Organizing Maps (SOM) on Microarray Gene Expressions

We refer the reader to a complete discussion of Self-Organizing Maps (SOM) as was presented in our WMSCI 2012 paper Lu and Segall ((2012), [15]) and we are thus providing below a brief discussion.

Self-Organizing Maps (SOM) belong to competitive neural networks. Competitive learning is an adaptive process in which neurons in a neural network are sensitive to different input categories, sets of samples in a specific domain of the input space. ((1), [7], [8], [10], [12], [13], [19], [20], [21], [32])

According to Wikipedia ((2013)[33]), a self-organizing map consists of components called nodes or neurons. Associated with each node is a weight vector of the same dimension as the input data vectors and a position in the map space. The self-organizing map describes a mapping from a higher dimensional input space to a lower
dimensional space. The procedure for placing a vector from data space onto the map is to find the node with the closest (smallest distance metric) weight vector to the data space vector.

A Self-Organizing Map consists of two layers as shown in figure 2. Suppose that we have a set of n-dimensional vectors. The first layer of SOMs is the input data which transfer to the second layer. The second layer has a number of neurons which are chosen arbitrarily and can be used to representing the feature space.

![SOMs Architecture](image)

**Figure 2: SOMs Architecture**

On the second layer, each neuron has the same dimension as the input neuron from the first layer. First of all, weights of the neurons on the second layer are set randomly. During the training process, they have their own weights vector and update those during the training process. When an input \( x \) arrives from the first layer to the second layer, the neuron that is best able to represent it wins the competition and is allowed to learn it even better. Moreover, not only the winning neuron but also its neighbors on the lattice are allowed to learn.

4. LYMPHOMA MICROARRAY GENE EXPRESSION PROFILE CLUSTERING

4.1 Background

After multi-agent chemotherapy, two subgroups of diffuse large-B-cell lymphoma had different outcomes. The germinal-center B-cell-like subgroup expressed genes that are characteristic of normal germinal-center B cell were associated with a good outcome. Whereas the activated B-cell-like subgroup expressed genes that are characteristic of activated blood B-cells were associated with a poor outcome. The international prognostic index (IPI) was generally used to stratify patients for therapeutic trials, but, its accuracy is not good enough.

In this paper, we explain how to check patients’ genes with microarrays and analyze for genetic abnormalities; find patients with distinctive gene expression profiles; and construct molecular predictors by using genes. There were 160 patients in the training set and 80 patients in the test set. The following three gene expression subgroups were identified: (i.) germinal center B-cell-like, (ii.) activated B-cell-like, and (iii.) type 3 diffuse large-B-cell lymphoma, but only the germinal center B-cell-like subgroup contributed to the lymphoma. Seventeen genes were used to construct a predictor of the survival after chemotherapy. Patients of the germinal center B-cell-like subgroup had the highest survival rate. We compared the accuracy of this predictor with that of the international prognostic index. By using data mining methods to analyze microarray gene expression data, we can create predictors for the survival after chemotherapy.

4.2 Experiments

For hierarchical clustering, we used correlation as similarity measure. We did complete linkage clustering of the 74 significant genes which distinguished between germinal center B-cell lymphoma and activated B-cell lymphoma, as shown in figure 3 and did single linkage clustering of all genes as shown in figure 4. Output describing the meaning of the each node on the hierarchical structure of the 74 significant genes has also been generated.

![Visualization of 7399 genes from 275 patient cases](image)

**Figure 3: Visualization of 7399 genes from 275 patient cases**

![The hierarchical structure of the 74 significant genes which can distinguish germinal center B-cell lymphoma and activated B-cell lymphoma](image)

**Figure 4: The hierarchical structure of the 74 significant genes which can distinguish germinal center B-cell lymphoma and activated B-cell lymphoma.**
5. MICROARRAY GENE EXPRESSION DATA CLASSIFICATION

We used 240 patient cases and 522 significant genes chosen by using t-test (p < 0.01). Three data mining algorithms are tested, which are Naïve Bayesian model, Random Forest model and Self Organizing Map. The experimental results are listed below where TP=True Positive and FP=False Positive.

<table>
<thead>
<tr>
<th>Table 1: Evaluation of Three Data Mining Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP Rate</td>
</tr>
<tr>
<td>Naive Bayesian</td>
</tr>
<tr>
<td>Random Forest</td>
</tr>
<tr>
<td>SOM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Statistics of Three Data Mining Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation Measurement</td>
</tr>
<tr>
<td>Correctly Classified Instances</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
</tr>
<tr>
<td>Kappa statistic</td>
</tr>
<tr>
<td>Mean absolute error</td>
</tr>
<tr>
<td>Root mean squared error</td>
</tr>
<tr>
<td>Relative absolute error</td>
</tr>
<tr>
<td>Root relative squared error</td>
</tr>
<tr>
<td>Total Number of Instances</td>
</tr>
</tbody>
</table>

Figure 5: Comparison of the precision and recall on naïve Bayesian, random forest and SOM models.

Data Analysis on Micorarray Lymphoma Gene Expression Data with 522 Genes from 240 Patient Cases

<table>
<thead>
<tr>
<th>Table 3: Microarray Significant Features in Gene Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Signature</td>
</tr>
<tr>
<td>Germinal-Center-B</td>
</tr>
<tr>
<td>Lymph-Node</td>
</tr>
<tr>
<td>MHC-Class II</td>
</tr>
<tr>
<td>Proliferation</td>
</tr>
</tbody>
</table>

With the microarray significant features which we used to make predictor, we checked with the gene signatures for germineral center B-cell signature, lymph-node signature, MHC-Class II signature and proliferation signature, which we used to make predictions in clinical practice. We can see 59.46% MHC-Class II signature characteristics are microarray significant features, 21.61% proliferation signature characteristics are microarray significant features, 3.64% lymph-Node signature characteristics are microarray significant features, and 2.65% germineral center B-cell signature characteristics are microarray significant features, which means microarray gene expression profiling and particular clinic pathological features are consistent. Therefore, there were 87.36% of microarray significant features in gene signatures and we can conclude that we can use microarray gene expression profiling alone to theoretically predict lymphoma.
6. DATA MINING OF MICROARRAY DATA FOR METASTASIS CANCER

6.1 Background

The data sets selected from the Broad Institute are two of those posted as available with unrestricted access as one of the web links posted on the web page of the Broad Institute Cancer Program Data Sets ((2008),(4)) and is that which is related to the “Genomic analysis of metastasis reveals an essential role for RhoC” research project of the Broad Institute. The selected data base for this research was used by Clark et al. ((2000),(5)) to illustrate the essential role of RhoC that is a member of the Rho family of proteins that promote reorganization of the cytoskeleton and regulate cell shape, attachment, and motility. Figure 6 from Wikepedia ((2008),(33)) provides an illustration of RhoC also known as “Ras homolog gene family, member C”. According to Wikepedia ((2008),(33)), overexpression of this gene is associated with tumor cell proliferation and metastasis.

6.2 Experiments

The databases utilized for this research in the applications of data mining are those used in Clark et al. ((2000),(5)) as collected at affiliated sites of the Broad Institute ((2008),(3)). These data was collected from human A375 tumor cells, and successive metastases M1, M2 and M3 that were isolated, expanded in tissue culture, and re-introduced into host mice which exhibited more pulmonary metastases. That is M2 data is that collected from those injected with A375M1 cells, and M3 data is that collected from those injected with A375M2 cells. These constitute the first set of data for which data mining had been subjected. The second data collection was for metastatic A375SM cells grown as a subcutaneous tumor to indicate that the expression of genes is truly intrinsic to the subjected metastatic cells. It was noted by the Broad Institute ((2008),(4)) that the tumor microenvironment may help to regulate the absolute level of gene expression.

The following figures were conducted using SAS Enterprise Miner version 5 using the data from Broad Institute ((2008),(4)) for A375 and A375SM tumor cells of metastasis cancer. Figure 7 and Figure 9 are the self-organized maps showing their frequency and normalized means, Figures 8 and 10 for the cluster proximities, and Tables 4 and 5 are the statistics from the SOM data mining. As can be seen from Figures 8 and 10, that the cluster proximities are generally much smaller of A375SM cells grown as a subcutaneous tumor. Figure 9 shows that the normalized means for A375SM are fewer but more intervals of frequency than those of Figure 6 for A375. Table 4 for A375 cells shows that the magnitude of the statistics are larger for those of the same segments of those of Table 5 for A375SM, indicating that the genes of the subcutaneous tumor are substantially and uniquely different from those of A375 cells of metastasis cancer.

Figure 6: RhoC Genome [Source: Wikipedia ((2013),(33))

Figure 7: Self-Organized Map for A375 tumor cells

Figure 8: Clusters Proximities for A375 tumor cells
This paper discussed an open issue in Microarray gene expression application. For the lymphoma study two hierarchical structures of microarray gene expression data were built with 7399 genes and 74 significant genes, that visualized the characteristics of the microarray gene expression profiles. We used naïve Bayesian model, random forest model and self organizing maps (SOM) to predict lymphoma with microarray gene expression profile from 240 patients. The experimental results showed us that lymphoma can be predicted with microarray gene expression data by using naïve Bayesian, random forest and SOM algorithms. We further compared the difference between clinic pathological features and microarray features by using gene signatures for Germinal center B-cell like lymphoma, lymph-node lymphoma, proliferation lymphoma and MHC-Class II lymphoma. We can conclude that, since clinical features and microarray features are associated with each other, the predictions from both clinic pathological features and microarray gene features are consistent.

For the metastasis cancer study we concluded that data mining of the microarrays using SOM was effective in distinguishing the uniqueness of the genes of the subcutaneous tumors.

**REFERENCES**


[4.] Broad Institute Cancer Program Data Sets (2008),

Table 5: Statistics from SOM Data Mining of A375SM Tumor Cells
www.broad.mit.edu/cancer/datasets.cgi


[18.] Lu, S., T. Hahn and R.S. Segall, “Data Mining on Survival Prediction after Chemotherapy for Diffuse Large B-Cell Lymphoma”, Poster presented at Joint Meeting of the Southern Section of the American Society of Plant Biologists (ASPB) and P3 Symposium, Little Rock, AR, April 6-9, 2013.


