

Robust Diagnosis of Non-Hodgkin Lymphoma Phenotypes Validated on Gene Expression Data from Different Laboratories

Gyan Bhanot^{1,3}
gyan@us.ibm.com

Gabriela Alexe^{1,3}
galex@us.ibm.com

Arnold J. Levine^{1,2}
ajlevine@ias.edu

Gustavo Stolovitzky³
gustavo@us.ibm.com

¹ Center for Systems Biology, Institute for Advanced Study, Princeton, New Jersey 08540, USA

² Robert Wood Johnson School of Medicine and Dentistry, Cancer Institute of New Jersey, New Brunswick, New Jersey 08903, USA

³ IBM Computational Biology Center, IBM Research, Yorktown Heights, New York 10598, USA

Abstract

A major challenge in cancer diagnosis from microarray data is the need for robust, accurate, classification models which are independent of the analysis techniques used and can combine data from different laboratories. We propose such a classification scheme originally developed for phenotype identification from mass spectrometry data. The method uses a robust multivariate gene selection procedure and combines the results of several machine learning tools trained on raw and pattern data to produce an accurate meta-classifier. We illustrate and validate our method by applying it to gene expression datasets: the oligonucleotide HuGeneFL microarray dataset of Shipp *et al.* (www.genome.wi.mit.edu/MPR/lymphoma) and the Hu95Av2 Affymetrix dataset (DallaFavera's laboratory, Columbia University). Our pattern-based meta-classification technique achieves higher predictive accuracies than each of the individual classifiers, is robust against data perturbations and provides subsets of related predictive genes. Our techniques predict that combinations of some genes in the p53 pathway are highly predictive of phenotype. In particular, we find that in 80% of DLBCL cases the mRNA level of at least one of the three genes p53, PLK1 and CDK2 is elevated, while in 80% of FL cases, the mRNA level of at most one of them is elevated.

Keywords: meta-classifier, feature selection, pattern, p53, gene expression, lymphoma

1 Introduction

The rapid development of microarray technologies [7, 16] allows the analysis of gene expression patterns to identify subsets of genes which are differentially expressed in different phenotypes (e.g., different types of cancer) and create personalized models for diagnosis and prognosis. There is a lot of ongoing research in developing tools and methodologies to extract information from biomedical data (e.g., [4, 6, 30]). However, there remains a need for a framework that can integrate the data from different laboratories and predictions from different techniques into a robust, noise insensitive predictive tool.

The aim of this study is to present such a tool, developed recently for cancer detection from SELDI-TOF mass spectrometry data [5], and adapt it for cancer diagnosis from gene expression data. We first apply a pattern-based multivariate approach to identify a subset of predictive genes out of a large pool of genes by requiring them to satisfy stringent filtering criteria. Next, we combine the predictions of several machine learning tools trained on the subset of predictive genes and on pattern data with the

aim of producing an accurate predictor. It is well-known [20, 26] that combining individual classifiers into a meta-classifier improves the error rate. In our method this effect is enhanced by using “pattern data,” which is a structured representation of the original data in a space in which patterns are viewed as synthetic variables.

We demonstrate our approach by creating a diagnostic model to distinguish follicular lymphoma (FL) from diffuse large B-cell lymphoma (DLBCL). Follicular lymphomas are one of the more common low grade non-Hodgkin’s lymphomas, which affect mostly adults, particularly the elderly [35]. They are of B-cell lymphocyte origin. Most cases of follicular lymphoma, especially those rich in small-cleaved cells, have a t(14;18) gene translocation, which results in a rearrangement and over-expression of the anti-apoptotic gene BCL-2. DLBCL is an aggressive form of non-Hodgkin lymphoma to which 25-60% FLs evolve over time. The FL transformation to DLBCL is associated with genetic alterations of p53 [23], p16 [25], p38MAPK [9], c-myc [15], BCL-6 [14]. Besides the genetic link, non-Hodgkin’ lymphomas could be caused by chemo and radiation therapy, and may also arise due to infections with the Epstein-Barr virus and HIV.

We will use the oligonucleotide microarray gene expression data of Shipp *et al.* [29] produced at the Whitehead Institute (WI data), and validate our findings on a separate Affymetrix gene expression data produced by DallaFavera laboratory at Columbia University (CU data, see [32]). The WI and CU datasets report gene expression data for DLBCL and FL cases which were obtained by using different Affymetrix chips (HuGeneFL chip for WI dataset and Hu95Av2 for the CU dataset). We also show that one can combine the two datasets into a single meta-dataset, while maintaining the accuracy of predictions.

To address the problem of differential diagnosis between FL and DLBCL from the WI data, Shipp and coworkers [29] used a signal-to-noise (SNR) correlation-based method to identify a subset of predictive genes and constructed a weighted-voting predictor based on the top 50 SNR correlated genes; their findings were validated through a leave-one-out cross-validation scheme on the same set of samples, and they obtained a sensitivity of 89% and a specificity of 100% for distinguishing FL from DLBCL cases. By applying different criteria (a multivariate pattern-based approach from Genes@Work [12] and a *t*-test [32] identified two additional subsets of predictive genes in the WI data. Stolovitzky further showed that about 88% of the genes in the union of his two subsets with the subset identified by Shipp *et al.* [29] have a consistent behavior in the independent CU data (i.e. are up-regulated or down-regulated in DLBCL vs. FL, respectively for both datasets).

Using our meta-classification method on a training subset of the WI data, we identify a robust subset of 30 predictive genes and construct a meta-classifier which misclassifies only one FL case when validated on the test set of the WI data and misclassifies only two FL cases when validated on the external CU data. We obtain biological insight by focusing on the subset of p53 responsive genes and extracted relevant patterns characteristic of FL and DLBCL. Finally, we illustrate how noisy results might be combined into a better predictive tool.

2 System and Methods

2.1 Datasets

The WI dataset [36] has 58 DLBCL samples and 19 FL samples. The data was obtained by using Affymetrix oligonucleotide microarrays (HuGeneFL chips) containing probes for 6817 genes. The CU dataset (DallaFavera laboratory, Columbia University) has 14 DLBCL and 7 FL samples obtained on Affymetrix microarrays (Hu95Av2 chips) containing probes for 12581 genes. In our study, DLBCL cases are referred to as positive, and FL cases as negative.

2.2 Patterns

A central concept involved in our method is that of a “pattern” in a two class dataset [3, 8]. In our approach, a positive pattern is a set of bounding conditions imposed on the intensity level of certain features (genes) which are satisfied by significantly many positive cases and by significantly few negative cases. A negative pattern is defined in a similar way.

Patterns are characterized by several parameters: (1) the *degree* of a pattern is the number of genes used in its defining conditions; (2) the positive (negative) *prevalence* of a pattern is the percentage of positive (negative) cases satisfying the defining conditions of the pattern; (3) the positive (negative) *homogeneity* of a pattern is the percentage of positive (negative) cases among all the cases satisfying the defining conditions of the pattern. High quality positive patterns have low degrees, and high positive prevalences and homogeneities. Figure 1 gives an example of positive and negative patterns in a two gene subspace.

One can generate a large number of patterns on a given dataset. Each pattern can be interpreted as a synthetic 0-1 variable associated with the samples in the dataset, the value 1 being assigned when the corresponding sample satisfies the defining conditions of the pattern, and the value 0 otherwise. Each sample is then represented by a vector with 0-1 entries, where each entry corresponds to a pattern. In this way, the original data can be represented in an abstract space which we call “pattern data” (see Figure 2). Since multiple patterns taken together define a region in the feature space where most of the samples of a given phenotype are located, pattern data provides additional structural information about the phenotype and can be used as input into machine learning algorithms to do phenotype recognition. Patterns can be extracted in an exhaustive way using a combinatorial algorithm described in [3].

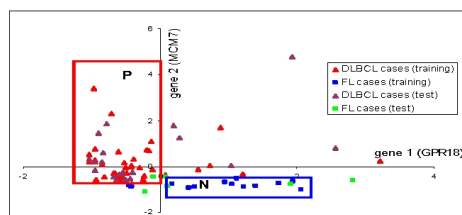


Figure 1: Examples of a positive pattern (P) and of a negative pattern (N).

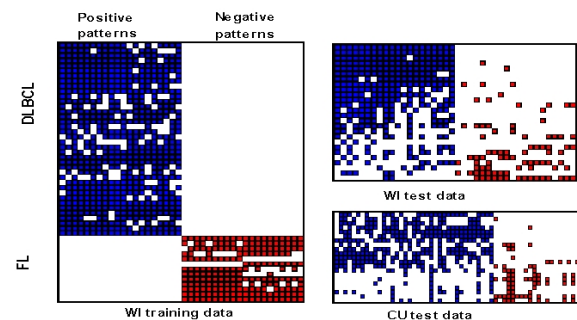


Figure 2: Visualization of training and test sets representation as pattern data: Each row corresponds to a case (DLBCL - positive or FL - negative) and each column corresponds to a pattern (positive or negative). A positive (negative) pattern is represented by a blue (or red) dot. Notice that in the training data DLBCL cases satisfy only positive patterns, and FL cases satisfy only negative patterns.

2.3 Overall Methodology

Our approach consists of four steps as described below and depicted in Figure 3:

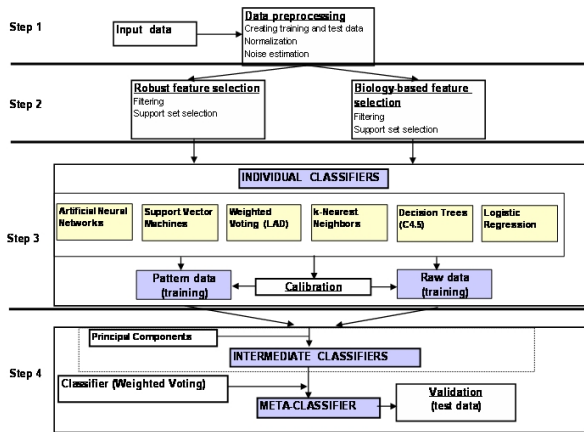


Figure 3: Flow chart of the meta-classifier approach.

2.3.1 Step 1: Data Preprocessing

From each dataset we selected only the genes that had P (present) calls in at least 50% of the samples. Then, following the normalization guidelines from the previous studies [29, 32], we set an upper limit of 16000 units and a lower ceiling of 20 units for all gene expression levels. The WI input data was 2/1 stratified sampled into a training and a test dataset. The CU data was considered as an external test data. Based on the assumption that a majority of genes are not differentially expressed across the FL and the DLBCL cases [19], the experimental noise was estimated from the data as normally distributed with mean 0 and variance equal to the median of the variances of all the genes across samples. For each array, the expression data was normalized by replacing the intensity level x of each gene g with $(x - \text{mean}(g)) / \sigma(g)$, where $\text{mean}(g)$ and $\sigma(g)$ represent the mean and the standard deviation of the intensity level of g across the samples in the training dataset.

2.3.2 Step 2: Finding Relevant Genes

In this step we tried to find a relatively small subset of genes which allow the construction of accurate predictive models. Numerous studies [11] assert that individual expression levels of genes are not informative enough for cancer classification and that combinations of expression levels of several genes might be more predictive. Based on this argument, we applied a two-step procedure [2]: in the first step we used a robust filtering approach to select a pool of features. In the second step we selected out of this pool a smaller set based on the significance of the genes in patterns (see below).

Filtering. We first selected those genes which showed a high correlation with phenotype. This filter was applied to the original data and to 500 other datasets, out of which 300 were obtained by perturbing the training data with experimental noise (normally distributed as $N(0, \lambda)$, where λ varies between 0.1 and 1), and 200 were obtained by perturbing its sample composition; the perturbation of the sample composition was performed by k -folding (randomly dividing the training data into k stratified parts and retaining in turn only $k-1$ out of the k parts), $k = 3, 5, 10$, and jackknifing (retaining in turn only 50 out of the 51 samples in the training data).

Two filtering options were considered: signal-to-noise and t -test with Benjamini-Hochberg false discovery rate correction for multiple selection. The choice between them was done by computing a variant of the robustness index R (introduced in [31]) for each of these two tests, and retaining the test with the highest index.

The signal-to-noise (SRN) score identifies the difference in means (the signal) in each of the two classes, i.e., the FL and DLBCL groups, scaled by the sum of the standard deviations (the noise): SRN

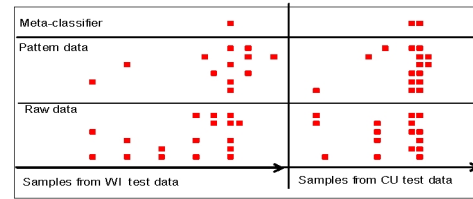


Figure 4: Errors of the meta-classifier compared to the individual classifier errors. The individual classifiers correspond to the rows of Table 3.

$= (\mu_1 - \mu_2)/(\sigma_1 + \sigma_2)$, where μ_i, σ_i are the mean and standard deviation of class i . The t statistics, defined as $(\mu_1 - \mu_2)/(\sigma_1^2 + \sigma_2^2)^{1/2}$, is an alternative form of the signal-to-noise statistics.

The robustness index R reflects the stability of the set of genes to filtering selection criterion when data is perturbed. In our approach R is defined as the proportion of genes which are selected by the filtering test in at least 90% of the 500 perturbed datasets. For an “ideal” filtering procedure R should be close to 1. At the other extreme, when the filtering procedure is reduced to the random selection of n genes, for sufficiently large n , one expects R to be close to zero. We verified by simulation that $R \sim 0.02$ for 500 random selections of 30 genes out of 2055.

Support set selection Next we selected a small set of genes from the filtered pool by studying the way in which combinations of these genes occur in patterns.

A collection of patterns was extracted from the filtered pool of genes by applying the combinatorial algorithm described in [3]. The optimal characteristic parameters of the patterns were determined by estimating the accuracy of a weighted-voting model constructed on pattern data through 10-fold cross-validation experiments on the training set, and by choosing those parameters for which the estimated accuracy was maximal. Out of the collection of patterns satisfying the optimal parameters we selected minimal subsets of patterns such that each case in the training data satisfies the defining conditions of at least 10 patterns. The features were ranked based on their frequency in the selected collection of patterns, where the prevalence of a pattern was taken as weight. At the end of this stage in the analysis, a relatively small set of top 10% of high ranked features was retained as a support set.

An alternative method of gene selection is to use relevant biological pathway information (biology-based gene selection). Numerous studies e.g., [13, 28] have noticed a correlation between over-expression of p53 and FL progression to DLBCL, and also that mutations of p53 are associated with histologic transformation in approximately 25% to 30% of FL cases. Other studies [22, 23] suggest that over-expression of MDM2 (and p53) identifies DLBCL and FL cases with poor prognosis, presumably because of alterations in the feedback loop between p53 and MDM2. We therefore focused our attention on the family of p53 regulated genes [10, 27] since we expect them to provide a robust signal. Our goal was to identify a subset of p53 responsive genes which, individually or in combinations, might be most predictive. The genes in the p53 set were pared down to 90 using the filtering and pattern selection criteria described above on the WI dataset. For this biology-based classification, we used the CU data as test data.

2.3.3 Step 3: Multiple Classifier Construction

Several different individual classifiers: artificial neural networks (ANN), support vector machines (SVM), weighted voting systems (WV), k -nearest neighbors (k NN), decision trees (C4.5) and logistic regression (LR) were trained and calibrated through cross-validation experiments on the raw and on the pattern training data as described in [5].

For this study we used the implementation of ANN, SVM, decision trees and LR provided in Weka [38], the implementation of WV provided in GenePattern [37] and the implementation of k NN provided in Genes@Work [39]. The classification accuracy of each individual classifier was estimated on the training data through a leave-one-out cross-validation experiment. The feature selection procedure was performed independently in each cross-validation experiment.

2.3.4 Step 4: Meta-Classifer Construction and Validation

The meta-classifier was defined as a weighted combination of the individual classifiers. The weight w_i of the classifier C_i is defined as $w_i = v_i/\|v\|$, where $v_i = \max[0, (\text{specificity}_i - 50\%)]\max[0, (\text{sensitivity}_i - 50\%)]$, $\|v\|$ is the L^1 norm of the vector v having the components v_i , and specificity_i and sensitivity_i are the specificity and the sensitivity of the classifier C_i obtained on leave-one-out cross-validation experiments on the training set. The meta-classifier prediction was then given by $P = \sum_i C_i w_i$, where $C_i = 1$ for DLBCL cases and -1 for FL cases. The meta-classifier will predict DLBCL with confidence

P if $P > 0$, and FL with confidence $|P|$ if $P < 0$. To increase the robustness of the meta-classifier with respect to the individual predictors, we imposed a threshold p for the certainty of the classification. The threshold p was computed on the training set as a p -value associated to the accuracy of the meta-classifier with respect to permutations of the sample class. Thus, a case was classified as DLBCL (FL) if $P > p$ (or $P < -p$). If $|P| < p$ the classification was considered “uncertain”. In real situations, the “uncertain” cases would be asked to repeat the tests to verify the results.

To validate the meta-classifier predictions we applied it to each sample in the WI test dataset and to each sample in the CU dataset. Furthermore, as in [5], we tested the robustness of the meta-classifier by perturbing the WI training dataset with experimental noise and then comparing the changes occurring on the predictions on the test set.

3 Results

3.1 Data Preprocessing

The WI dataset was split into a training set consisting of 51 samples (38 DLBCL and 13 FL cases), and a test set consisting of 20 DLBCL and 6 FL cases. After ceiling and normalization, we eliminated the genes with no variation across the samples in the WI data and in the CU data, respectively. The 50% call filtering criterion was passed by only 2055 of the 6817 genes in the WI dataset. Of these, only 1901 passed the filtering criterion in the CU data. Thus the WI training and test sets for analysis had 2055 genes, and the CU external test had 1901 genes.

3.2 Finding Relevant Genes

Filtering. The robustness indices of the signal to noise correlation and t -test were $R=0.27$ and 0.23 , respectively. Hence, the signal-to-noise correlation was chosen as the filtering criterion. A pool of 73 genes passed the top 25% signal-to-noise test in at least 90% of the perturbed datasets. The selected pool contained 51 of the 100 top genes selected by [29], 32 of the 100 genes selected by the method in Genes@Work [32], 28 genes selected based on the t -statistics [32], and only 9 p53 responsive genes.

Support set selection. A collection of 1595 positive and 667 negative patterns of degree 2 with positive (or negative) prevalence above 50%, were extracted from the restriction of the training dataset to the pool of 73 genes. Out of this collection we selected a subset of 57 (37 positive and 20 negative) patterns based on the criterion that each case in the training data satisfies at least 10 of the selected patterns. Table 1 presents the collection of 30 genes that occurred in the definition of the selected 57 patterns.

Pattern data was defined with respect to the 57 selected patterns and can be visualized in Figure 2. To illustrate how pattern data provides structural information about the cases, Table 2 presents some examples of patterns (combinations of conditions) which are characteristic of large subgroups of DLBCL and FL cases. The striking feature of Table 2 is the fact that simple conditions on a few genes are able to generate a very clean classification of the training data and an accurate prediction on the test data.

3.3 Meta-Classifier Construction and Validation

We trained 6 individual classifiers (see Figure 3) on raw and pattern training data using the 30 robust genes and assessed their performance on the training data through leave-one-out cross validation experiments. We found that the error distribution of the individual classifiers on the training was uncorrelated, with only one false positive error for which 33% of the predictors agreed. We noticed that the average performance of the individual classifiers was better on the pattern data (average sensitivity 100% and average specificity 96.2%) than on the raw data (average sensitivity 95.6% and average specificity 91.0%) on the training set. Weighted voting was the best individual classifier, and

Table 1: Support set of 30 robust genes sorted in decreasing order of their signal-to-noise score. The top 16 genes are up-regulated in FL cases.

Gene symbol	Shipp et al.	Genes@Work	t-test	p53 regulated	Biological function
SEPP1	*	*	*		oxidative stress
TXNIP	*	*			metastases suppressor
DNASE1L3	*	*			apoptosis
CDH11	*	*	*		cell adhesion
LUCA15	*	*			apoptosis
GPR18	*	*	*		signaling pathway
CLU	*	*	*		apoptosis
LY9	*	*			cell adhesion
RHOH	*	*			T-cell differentiation
ELF2					transcription
CCNG2				*	cell cycle
CR2					complement activation
CDKN2D				*	cell cycle
PPP2R5C		*			signal transduction
G18					cell growth
LY86		*			apoptosis
ARPC1B					cell motility
MCM7	*	*	*	*	cell cycle
BCL2A1	*	*	*		apoptosis
IMPDH2	*	*	*		GMP biosynthesis
RRP45	*				immune response
STAT1					NF-kappaB cascade
DLG7	*	*	*		cell-cell signaling
SLC1A5	*	*	*		transport
TUBB2	*	*	*		microtubule movement
PSMA6					protein catabolism
PSMC1	*	*	*		spinocerebellar ataxia
LGALS3	*	*	*		sugar binding
CLTA	*	*	*		transport
PAGA	*	*	*		cell proliferation

Table 2: Examples of characteristic patterns for DLBCL and FL

Pattern	Gene symbol								Prevalence (%)			
	TXNIP	CDH11	GPR18	ELF2	MCM7	STAT1	PSMA6	CLTA	Training set		Test set	
									Pos	Neg	Pos	Neg
P1		≤ 0.46			> 0.78				89	0	88	15
P2				≤ 0.28			> -0.75		87	0	71	8
N1			> -0.03			≤ -0.44			0	85	9	46
N2	> 0.14							≤ -0.46	0	85	0	23

Table 3: Performance of classifiers on training data (validated through leave-one-out experiments) and on the test data (validated directly).

Classifier	Weight	Training			Test			
		Sensitivity (%)	Specificity (%)	Error rate (%)	Sensitivity (%)	Specificity (%)	Error rate (%)	
Trained on raw data	ANN	0.08	94.74	92.31	5.88	82.35	84.62	17.02
	SVM	0.08	97.37	92.31	3.92	97.06	76.92	8.51
	kNN	0.09	97.37	100.00	1.96	91.18	84.62	10.64
	WV	0.07	92.11	92.31	7.84	94.12	76.92	10.64
	C4.5	0.06	94.74	84.62	7.84	94.12	69.23	12.77
Trained on pattern data	LR	0.07	97.37	84.62	5.88	94.12	69.23	12.77
	ANN	0.10	100.00	100.00	0.00	97.06	76.92	8.51
	SVM	0.10	100.00	100.00	0.00	97.06	76.92	8.51
	kNN	0.10	100.00	100.00	0.00	100.00	69.23	8.51
	WV	0.10	100.00	100.00	0.00	97.06	76.92	8.51
C4.5	0.10	100.00	100.00	0.00	91.18	76.92	12.77	
LR	0.05	100.00	76.92	5.88	100.00	61.54	10.64	
Meta-classifier		100.00	100.00	0.00	100.00	76.92	6.38	

logistic regression the worst. Except for logistic regression, all the individual classifiers performed with 100% accuracy on the pattern data (see Table 3).

We constructed the meta-classifier as a weighted combination of the individual classifiers. Figure 4 presents the predictions of the individual classifiers and of the meta-classifier on the test dataset. Notice that the predictions of the meta-classifier are better than the predictions of any individual classifier.

We further perturbed the raw training data with experimental noise, generated the corresponding pattern data, re-trained the classifiers and constructed the meta-classifier on the perturbed training dataset. The error distributions of the individual classifiers and of the meta-classifiers had only a small variance and in fact the meta-classifier made the same errors as on the non-perturbed data. The fact that the meta-classifier predictions did not change is a confirmation of its robustness, particularly of its stability to experimental noise.

3.4 Biology-Based Gene Selection: Role of p53 Regulated Genes

The p53 tumor suppressor gene is involved in pathways associated with cell division, hypoxia, apoptosis, cell signaling, DNA damage, DNA repair, etc. [27]. From the set of p53 associated genes, 215 were present in both WI and CU datasets. About 32% of the genes from this list were oppositely regulated in the CU and WI datasets, hence we discarded them as inconsistent. The 90 p53 responsive genes shown in Table 4 were able to differentiate between DLBCL and FL with a p -value below 0.01 with respect to the t -test in both the WI and CU data. Once again, the Benjamini-Hochberg false

discovery rate correction for multiple selection was applied.

The core regulators of p53 which were identified by our methods in the WI and CU datasets are MDM2 and E2F1. We found that the expression level of MDM2 is consistently up-regulated on the FL vs. DLBCL cases (t -test p -value 0.10), while the expression level of E2F1 is consistently down-regulated (t -test p -value 0.08). Figure 5 shows the pattern data constructed on the WI training data and its performance on the CU test data. A weighted voting classifier trained on patterns extracted from the 90 genes presented in Table 4 made the same 2 false positives on the CU test data as our meta-classifier, and one additional false negative (DLC14). If we have used 90 randomly selected genes from set of 215 genes, the average sensitivity would have been 97.72% (CI95% 95.62 - 99.81%) and the specificity would have been 7.14% (CI95% 6.90 - 21.14%) for 100 random experiments.

Table 4: List of top 90 p53 responsive genes in increasing order of their p -values (from 3×10^{-11} to 0.005). Genes up-regulated in the FL cases are marked with (*).

Gene symbol			
CCNB1	EPRS	TOPBP1	CDK7
MCM7	GSK3B	PMAIP1	E2F3
BRCA1	COL6A1	ACA2	MDM4
BCL2A1	HRAS	E2F5*	AMPD2
PPP2R4	SERPING1	POLA	RBBP4
EIF2S2	CCNA2	HMGB2	CCNG2*
COMT	CCT6A	PSMB5	HARS
IARS	MCM2	ACTA2	CASP6
MP1	PRKDC	INSR	RPS6KA1
ALAS1	CAD	SNRPA	GRP58
MRPL3	TNFRSF1B	G1P2	TP53
NCF2	ZNF184*	IMPDH1	SMAD2
AARS	ALDOA	MAP2K2	ATP5C1
KIF11	KARS	TOP2A	TIMP3
CDK4	MAD2L1	CXCL1	THBS2
ATP1B1	GOT1	BAG1	MYCBP
CDC20	CDC25B	TOP1	DTR
PRIM1	PSMA1	MAP4	TIMP3
CDC2	KIAA0101	FDFT1	CBS
TOP2A	PCNA	MTA1	CDKN2D*
CDK2	TCF3	CDKN1A	RELA
MYC	CYC1	HLAE*	
CCNE1	UPP1	PLK1	

Table 5: Examples of high prevalence p53 patterns.

Pattern	Gene symbol				Prevalence (%)			
	MCM7	BCL2A1	JUN	ZNF184	Training set		Test set	
					Pos	Neg	Pos	Neg
P1	≥ -0.77	≥ -0.92			97	21	86	29
P2	≥ -0.77		≥ -0.85		97	26	57	14
N1	≤ -0.61	≤ -0.28			2	84	0	57
N2		≤ -0.74		≥ -0.1	2	58	0	0

4 Discussion

In this study the WI gene expression data for DLBCL and FL [29] was re-analyzed using a pattern-based meta-classification technique and the results were validated on an external dataset (CU). Several other studies of the WI and CU data (e.g., [1, 9]) have reported a high accuracy in distinguishing the two diseases. Our motivation for reanalyzing the data was to exhibit a robust pattern-based meta-classification method for accurate predictions which has its roots in the pattern-based weighted voting classification method called Logical Analysis of Data [8].

Our method selected a robust subset of 30 genes with low sensitivity to experimental noise and sample composition. Only 19 of these 30 genes were selected by [29]. These 19 genes are known either to play an active role in cancer (e.g., BCL2A1, DLG7, MCM7) or to be involved in cell metabolism, cell growth, cell motility, cell adhesion etc. Of the remaining 11 genes in our 10 are up-regulated in FLs and only one (STAT1) is up-regulated in DLBCL. Moreover, 7 of these 11 genes (CDKN2D, CCNG2, RBM5, STAT1, G18, LY86, PPP2R5C) are well known to play a role in cancer (see e.g., [39]). The final phenotype prediction was obtained by integrating individual classifications into a meta-classifier. We noticed that the errors produced by the individual classifiers were not correlated, and the overall performance of the meta-classifier was superior to each individual classifier.

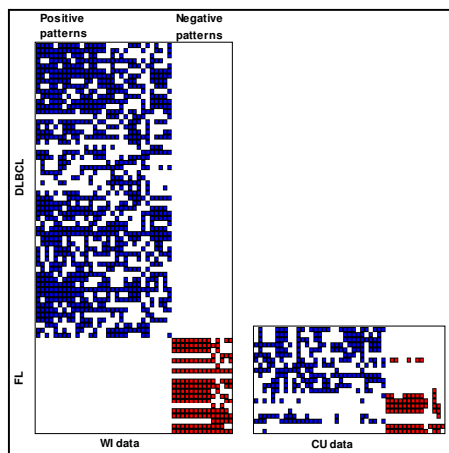


Figure 5: Visualization of p53-pattern data.

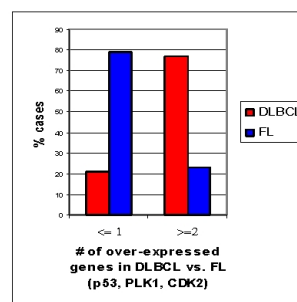


Figure 6: Histogram of % DLBCL (blue) and FL (red) cases having one or more of the three genes (p53, PLK1, CDK2) over-expressed.

The role of p53 responsive genes. Special attention was paid to the role of p53 responsive genes in differentiating between follicular and diffuse large cell lymphomas. We noticed that our input list of 215 p53 responsive genes does not contain several important p53 regulated genes which are known to respond to activated p53 in apoptosis (e.g., Pidd, Bax, Noxa, Puma, Siah, Perp, etc.), or in inhibition of angiogenesis and metastasis (e.g., Pai, Bai1, Kai, etc.). However, we found that several p53 genes in our list which are known to be involved in cell cycle arrest (Cyclin E, Cdk2, p21, Cyclin B, Cdc2) are up-regulated in DLBCLs'and down-regulated in FLs (p -value 0.01) in both WI and CU datasets. The genes involved in DNA repair (e.g., p48 and R2) are up-regulated for the DLBCL cases in the WI data, but they do not have consistent behavior in the CU data.

10 of the 90 p53 responsive genes presented in Table 4 were also selected by Genes@Work [12], 5 were selected by the t -test and by the signal to noise correlation criterion [29, 32], and 4 genes (MCM7, BC2A1, CDNK2D and CCNG2) were selected in our support set of 30 genes.

A new biomarker for non-Hodgkin lymphomas. We noticed in our data that p53 is consistently up-regulated (p -value 0.005) in the DLBCL vs. FL cases in both datasets, which is a confirmation of previous studies e.g. [28]. However, p53 alone can only differentiate DLBCL vs. FL cases with a sensitivity of 70% and a specificity of 50%, and so, by itself, it is not a very accurate biomarker. However, the combination of p53 with other 2 genes (PLK1 and CDK2) is an 80% accurate biomarker of the FL progression to DLBCL. At most one of the three genes is up-regulated in 77% of the FL cases and at least two of the three genes are up-regulated in 79% of the DLBCL cases. This is shown in Figure 6.

These findings are in agreement with recent results regarding PLK1 [33] where it is suggested that PLK1 is a potential target for cancer therapy and a new prognostic marker for cancer, and [21] showed that PLK1 is a potential biomarker for DLBCL.

Among other significant biomarkers we identified from p53 responsive genes we would mention MCM7, ZNF184, ALDOA, BCL2A1, CCNB1, MDM4 (for example, MCM7 alone is able to distinguish between DLBCL and FL with 79.49% accuracy on the WI data). However, combinations of the p53 responsive genes may have even more predictive value. In Table 5 we present some patterns for four p53 biomarkers which are good predictors.

Acknowledgments

The authors thank the anonymous referees for their interesting comments. GA was supported by the New Jersey Commission on Cancer Research (CCR-703054-03) and by The David and Lucile Packard

Foundation and The Shelby White and Leon Levy Initiative Fund. GB thanks IAS for visiting member status.

References

- [1] Alexe, G., Alexe, S., Axelrod, E.D., Hammer, P.L., and Weissmann, D., Logical analysis of diffuse large B cell lymphomas, *Artificial Intelligence in Medicine*, in press.
- [2] Alexe, G., Alexe, S., Hammer, P.L., and Vizvari, B., Pattern-based feature selection in genomics and proteomics, Rutgers University, *RUTCOR Research Report RRR*, 7:1–24, 2003 <http://rutcor.rutgers.edu/~rrr/2003.html>.
- [3] Alexe, G. and Hammer, P.L., Spanned patterns in Logical Analysis of Data, *Discr. Appl. Math.*, in press (also available as technical report at <http://rutcor.rutgers.edu/~rrr/2002.html>).
- [4] Armstrong, N.J. and van de Wiel, M.A., Microarray data analysis: from hypotheses to conclusions using gene expression data, *Cell Oncol.*, 26(5-6):279–290, 2004.
- [5] Bhanot, G., Alexe, G., Venkataraghavan, B., and Levine, A.J., A robust meta-classification strategy for cancer detection from mass spectrometry data, *Proteomics*, in press (<http://www.wiley-vch.de/publish/en/journals/alphabeticalIndex/2120/>).
- [6] Califano, A., Stolovitzky, G., and Tu, Y., Analysis of gene expression microarrays for phenotype classification. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:75–85, 2000.
- [7] Choudhuri, S., Microarrays in biology and medicine, *J. Biochem. Mol. Toxicol.* 18(4):171–179, 2004.
- [8] Crama, Y., Hammer, P.L., and Ibaraki, T., Cause-effect relationships and partially defined boolean functions, *Ann. Oper. Res.*, 16: 299–326, 1988.
- [9] Elenitoba-Johnson, K.S., Jenson, S.D., Abbott, R.T., Palais, R.A., Bohling, S.D., Lin, Z., Tripp, S., Shami, P.J., Wang, L.Y., Coupland, R.W., Buckstein, R., Perez-Ordenez, B., Perkins, S.L., Dube, I.D., and Lim, M.S., Involvement of multiple signaling pathways in follicular lymphoma transformation: p38-mitogen-activated protein kinase as a target for therapy, *Proc. Natl. Acad. Sci. USA*, 100(12):7259–7264, 2003.
- [10] Finlay, C.A., Hinds, P.W., and Levine, A.J., The p53 proto-oncogene can act as a suppressor of transformation, *Cell*, 57(7):1083–1093, 1989.
- [11] Guyon, I. and Elisseeff, A., An introduction to variable and feature selection, *J. Mach. Learn. Res.*, 3:1157–1182, 2003.
- [12] Lepre, J., Rice, J.J., Tu, Y., and Stolovitzky, G., Genes@Work: an efficient algorithm for pattern discovery and multivariate feature selection in gene expression data, *Bioinformatics*, 20(7):1033–1044, 2004.
- [13] Lo Coco, F., Gaidano, G., Louie, D.C., Offit, K., Chaganti, R.S., and Dalla-Favera, R., p53 mutations are associated with histologic transformation of follicular lymphoma, *Blood*, 82(8):2289–2295, 1993.
- [14] Lossos, I.S., Alizadeh, A.A., Diehn, M., Warnke, R., Thorstenson, Y., Oefner, P.J., Brown, P.O., Botstein, D., and Levy, R., Transformation of follicular lymphoma to diffuse large-cell lymphoma: alternative patterns with increased or decreased expression of c-myc and its regulated genes, *Proc. Natl. Acad. Sci. USA*, 99(13):8886–8891, 2002.

- [15] Lossos, I.S., Jones, C.D., Zehnder, J.L., and Levy, R., A polymorphism in the BCL-6 gene is associated with follicle center lymphoma, *Leuk. Lymphoma*, 42(6):1343–1350, 2001.
- [16] Lyons P., Advances in spotted microarray resources for expression profiling, *Brief. Funct. Genomic Proteomic*, 2(1):21–30, 2003.
- [17] Matolcsy, A., Warnke, R.A., and Knowles, D.M., Somatic mutations of the translocated bcl-2 gene are associated with morphologic transformation of follicular lymphoma to diffuse large-cell lymphoma, *Ann. Oncol.*, 8:119–122, 1997.
- [18] McDonnell, T.J., Deane, N., Platt, F.M., Nunez, G., Jaeger, U., McKearn, J.P., and Korsmeyer, S.J., Bcl-2-immunoglobulin transgenic mice demonstrate extended B cell survival and follicular lymphoproliferation, *Cell*, 57(1): 79–88, 1989.
- [19] McShane, L., Radmacher, R.D., Freidlin, B., Yu, R., Li, M.C., and Simon, R., Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data, *Bioinformatics*, 18:1462–1469, 2002.
- [20] Merz, C., *Classification and regression by combining models*, Dissertation, UCI, 1998.
- [21] Mito, K., Kashima, K., Kikuchi, H., Daa, T., Nakayama, I., and Yokoyama, S., Expression of Polo-Like Kinase (PLK1) in non-Hodgkin’s lymphomas, *Leuk. Lymphoma*, 46(2):225–231, 2005.
- [22] Moller, M.B., Nielsen, O., and Pedersen, N.T., Frequent alteration of MDM2 and p53 in the molecular progression of recurring non-Hodgkin’s lymphoma, *Histopathology*, 41(4):322–330, 2002.
- [23] Moller, M.B., Nielsen, O., and Pedersen, N.T., Oncoprotein MDM2 overexpression is associated with poor prognosis in distinct non-Hodgkin’s lymphoma entities, *Mod. Pathol.*, 12(11):1010–1016, 1999.
- [24] Monti, S., Tamayo, P., Mesirov, J., and Golub, T., Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data, *Machine Learning J.*, 52(1-2):91–118, 2003.
- [25] Pinyol, M., Cobo, F., Bea, S., Jares, P., Nayach, I., Fernandez, P.L., Montserrat, E., Cardesa, A., and Campo, E., p16(INK4a) gene inactivation by deletions, mutations, and hypermethylation is associated with transformed and aggressive variants of non-Hodgkin’s lymphomas, *Blood*, 91(8):2977–2984, 1998.
- [26] Prodromidis, A.L. and Stolfo Salvatore, J.A., Comparative evaluation of meta-learning strategies over large and distributed data sets, *Sixteenth International Conference on Machine Learning*, Bled Slovenia, 18–27, 1999.
- [27] Robins, H., Alexe, G., Harris, S., and Levine, A.J., The first twenty-five years of p53 research, *Cell*, in press.
- [28] Sander, C.A., Yano, T., Clark, H.M., Harris, C., Longo, D.L., Jaffe, E.S., and Raffeld, M., p53 mutation is associated with progression in follicular lymphomas, *Blood*, 82(7):1994–2004, 1993.
- [29] Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., Ray, T.S., Koval, M.A., Last, K.W., Norton, A., Lister, T.A., Mesirov, J., Neuberg, D.S., Lander, E.S., Aster, J.C., and Golub, T.R., Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, *Nature Med.*, 8(1):68–74, 2002.

- [30] Slonim, D.K., From patterns to pathways: gene expression data analysis comes of age, *Nat. Genet.*, 32:502–508, 2002.
- [31] Stolovitzky, G., Gene selection in microarray data: the elephant, the blind men and our algorithms, *Curr. Opin. Struct. Biol.*, 13(3):370–376, 2003.
- [32] Stolovitzky, G., Gene selection strategies in microarray expression data: applications to case-control studies, In Deisboeck T.S., Kresh J.Y., and Kepler T.B. (eds): *Complex Systems Science in BioMedicine*. Kluwer/Plenum Publishers, NY, in press (2005) (preprint: <http://www.wkap.nl/prod/a/Stolovitzky.pdf>).
- [33] Takai, N., Hamanaka, R., Yoshimatsu, J., and Miyakawa, I., Polo-like kinases (Plks) and cancer, *Oncogene*, 24(2):287–291, 2005.
- [34] Tu, Y., Stolovitzky, G., and Klein, U., Quantitative noise analysis for gene expression microarray experiments, *Proc. Natl. Acad. Sci. USA*, 99(22):14031–14036, 2002.
- [35] Winter, J.N., Gascoyne, R.D., and Van Besien, K., Low-grade lymphoma, *Hematology*, 203–220, 2004.
- [36] <http://www-genome.wi.mit.edu/MPR/lymphoma>
- [37] <http://www.broad.mit.edu/cancer/software/genepattern/>
- [38] <http://www.cs.waikato.ac.nz/~ml/weka/>
- [39] <http://www.infobiogen.fr/services/chromcancer/>
- [40] <http://www.research.ibm.com/FunGen>