BIOINFORMATICS

Vol. 1 no. 1 2003
Pages 1–1

Robust diagnosis of non-Hodgkin lymphoma phenotypes validated on gene expression data from different laboratories

Gabriela Alexe[1,], Gyan Bhanot[1,3], Arnold J. Levine[1,2] and Gustavo Stolovitzky[3]*

[1]*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

Motivation: One of the major challenges in cancer diagnosis from microarray data is to develop highly accurate and robust classification models which are independent of the analysis techniques used and can combine data from different laboratories.

Methods: We propose a novel, robust meta-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 distinguish diffuse large B-cell lymphoma (DLBCL) from follicular lymphoma (FL) on two independent gene expression datasets: the oligonucleotide HuGeneFL microarray dataset of Shipp et al. (www. genome.wi.mit.du/MPR/lymphoma) and the Hu95Av2 Affymetrix dataset (DallaFavera's laboratory, Columbia University).

Results: The pattern-based meta-classification technique achieves higher predictive accuracies than each of the individual classifiers trained on the same dataset, is robust against various data perturbations and provides subsets of predictive genes. We also find that combinations of p53 responsive genes are highly predictive of phenotype. In particular, we find that in DLBCL cases the mRNA level of *at least* one of the three genes p53, PLK1 and CDK2 is elevated, while in FL cases, the mRNA level of *at most* one of them is elevated.

* To whom correspondence should be addressed.

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

1 INTRODUCTION

The rapid development of microarray technologies (Choudhuri, 2004) (Lyons, 2003) allows the analysis of gene expression patterns to identify subsets of genes which are differentially expressed between different phenotypes (e.g., different types of cancer), and to integrate data into personalized models capable of providing diagnosis and predicting prognosis. There is a lot of ongoing research in developing tools and methodologies to extract information from biomedical data (e.g., (Califano et al., 2000), (Armstrong et al., 2004), (Slonim, 2002), (Wright et al., 2003)). 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, recently developed for cancer detection from SELDI-TOF mass spectrometry data (Bhanot et al., 2004), 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 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 (Merz, 1998) (Prodromidis et al., 1999) that combining individual classifiers into a meta-classifier has the effect of improving the error rate. In our method this effect is boosted 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 diagnosis model to accurately distinguish between follicular lymphoma (FL) and diffuse large B-cell lymphoma (DLBCL). We will use the oligonucleotide microarray gene expression data of Shipp et al. (2002) 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 (Stolovitzky, 2005)). 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 (2002) 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 (Lepre et al., 2004) and a t-test (Stolovitzky, 2005) 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. (2002) have a consistent behavior in the independent CU data (i.e.. are up-regulated or down-regulated in DLBCL vs. FL, respectively).

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

2 SYSTEM AND METHODS

Datasets

The WI dataset (http://www-genome.wi.mit.edu/MPR/ lymphoma) 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.

Follicular lymphomas are one of the more common low grade non-Hodgkin's lymphomas, which affect mostly adults, particularly the elderly (Winter et al., 2004). 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 antiapoptotic 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 (Moller et al., 1999), p16 (Pyniol, 1998), p38MAPK (Elenitoba-Johnson et al., 2003), c-myc (Lossos et al., 2002), BCL-6 (Lossos et al, 2001). 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.

Overall methodology

Our approach consists of the following steps (see Figure 1):

1. Data preprocessing: involves creating training and test data, data normalization, noise estimation.
2. Robust feature selection: involves a two-step procedure for extracting a robust subset of predictive genes and the creation of pattern data. It can also include a biology-based extraction of a subset of genes and the creation of corresponding pattern data.
3. Multiple classifiers construction: involves applying several classification methods on the raw and pattern training data and evaluating their performance by leave-one-out cross validation experiments on training data.
4. Meta-classifier construction and validation: involves combining the predictions of individual classifiers to generate a prediction of the phenotype. The accuracy of the resulting meta-classifier is tested on test data, not necessarily produced by the same laboratory.

Data preprocessing

From each dataset we selected only the genes that had present calls in at least 50% of the samples. Then, following the reasoning in a previous study (Shipp et al., 2002), we set an upper ceiling of 16000 units and a lower ceiling of 20 units for all gene expression levels. 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 mean($g$) and $\sigma(g)$ represent the mean and the standard deviation of the intensity level of g across the samples in the dataset.

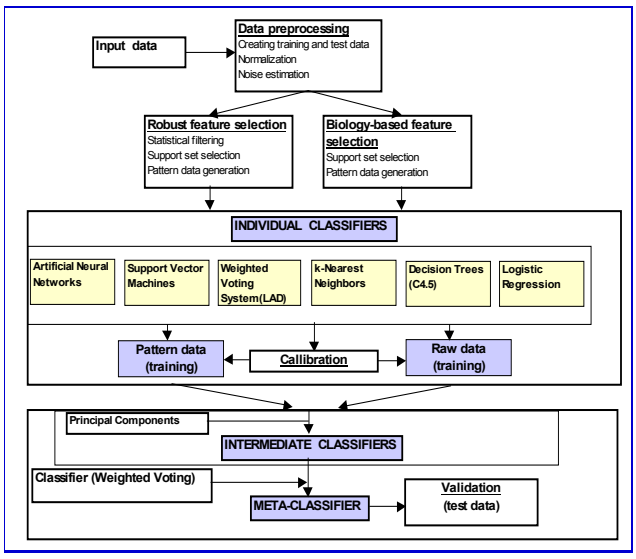

Fig. 1. Flow chart of the meta-classifier approach.

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 (McShane et al., 2002), 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.

Robust feature selection

We applied a two-step feature selection procedure similar to the wrapping approach presented in Alexe et al. (2003). We first applied a robust single gene filtering procedure which selected those genes which showed a relatively high (top 25%) signal-to-noise correlation with the phenotype. To ensure robustness to experimental noise and sample composition, we created 500 perturbed training datasets in two ways: (1) by adding different levels of experimental noise $N(0, \lambda)$, where λ varies between 0.1 and 1, and (2) by bootstrapping and jackknifing samples out of the training data (see also (Tu et al., 2003)). By applying the filtering approach to each of the perturbed training datasets, as well as to the original data, we selected a set of genes as the top 25% genes with respect to signal-to-noise correlation in at least 90% of the perturbed datasets. The filtering criterion was chosen based on a variant of the robustness index introduced in (Stolovitzky, 2003).

In the second step, we used a combinatorial pattern recognition algorithm (Alexe et al., in press) to extract large collections of high quality patterns (rules) from the training data restricted to the genes in the pool selected in the previous step.

A positive pattern is defined by a set of bounding conditions imposed on the intensity level of certain genes which are satisfied by significantly many positive cases and by significantly few negative cases. A negative pattern is defined in a similar way. This definition for patterns differs from that one used in (Lepre et al., 2004), where a pattern is defined as a subset of genes with a very low variation of their intensity level across significantly many positive cases in the data. 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 2 gives an example of positive and negative patterns in a 2 gene subspace.

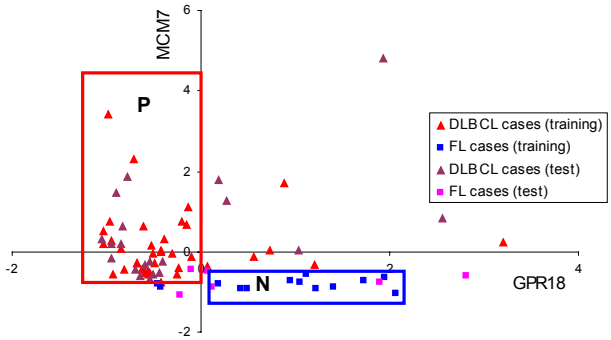

Figure 2. Examples of a positive pattern (P) and of a negative pattern (N).

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 3).

We determined the optimal characteristic parameters of the patterns by estimating the accuracy of a weighted-voting model constructed on pattern data through 10-fold cross-validation experiments on the training set, and chose those parameters for which the estimated accuracy was maximal, as detailed in (Bhanot et al., 2004). 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.

We define a support set as a set of predictive genes selected by a certain procedure (e.g., t-test etc). In our case, the support set was defined as a subset of pairwise low correlated genes occurring in the definition of the selected patterns.

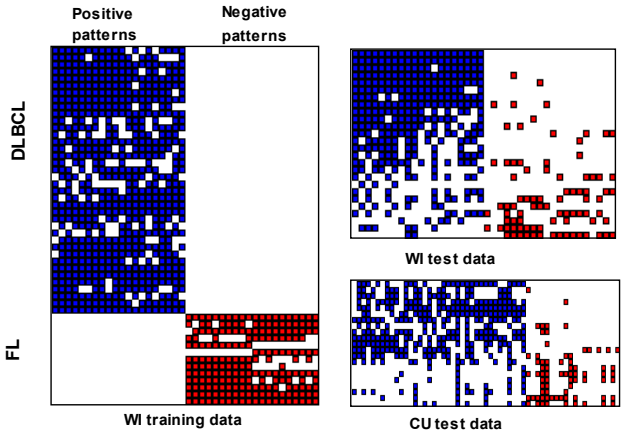

Figure 3. Visualization of training and test sets representation as pattern data: Each row corresponds to a case (DLBCL or FL) and each column corresponds to a pattern (positive or negative). A positive (negative) pattern is represented by a blue (or red) dot. Notice that in the training data DLBCL cases satisfy only positive patterns, and FL cases satisfy only negative patterns.

Multiple classifier construction

Several different individual classifiers: artificial neural networks (ANN), support vector machines (SVM), weighted voting systems (WV), k-nearest neighbors (kNN), 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 (Bhanot et al., 2004). For this study we used the implementation of ANN, SVM, decision trees and LR provided in Weka (http://www.cs.waikato.ac.nz/ ~ml/weka/) and the implementations of WV and kNN provided in GenePattern (http://www.broad.mit.edu/cancer/software/genepattern/) and Genes@Work (http://www.research.ibm.com/FunGen). The classification accuracy of each individual classifier was estimated on the training data through a leave-one-out cross-validation experiment.

Meta-classifier 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, ($specificity_i$ – 50%)]max[0,($sensitivity_i$ –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 $P = \sum_i Ci\ wi$, 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 need additional classifiers or improved subsets of predictive genes.

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 (Bhanot et al., 2004), 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.

Biology-based gene selection: Role of p53 regulated genes

Numerous studies e.g., (Sander et al., 1993), (LaCoco et al., 1993) 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 (Moller et al., 1999), (Moller et al., 2002) 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 (Finlay et al., 1989), (Robins et al., in press) 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.

Meta-data analysis: extracting information from multiple support sets

We tested the validity of the assumption that the information provided by different support sets identified in gene expression data might boost the predictive power of a classifier trained on the data. We created a meta-dataset by merging the WI and CU datasets and trained a weighted voting (WV) classifier on each of the support sets identified in previous studies (Shipp et al., 2002), (Stolovitzky, 2005), and in the current study. The predictions provided by each support set were weighted based on the WV performance on leave-one-out experiments and integrated into a novel meta-classifier.

3 RESULTS

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 out of the 6817 genes in the WI dataset. Of these genes, only 1901 passed the filtering criterion in the CU data. The WI training and test sets were described by the selected 2055 genes, and the CU external test set by the selected subset of 1901 genes.

Robust feature selection

Filtering. We generated 200 datasets by perturbing the training data with experimental noise N(0,λ), with λ = 0.1, 0.2, …, 1.0 (20 datasets for each value of λ). In addition, we generated 300 other datasets by perturbing the sample composition of the training data. The perturbation of the sample composition was performed by bootstrapping the samples (extracting with replacement n=51 samples from the training data and retaining the subset of those samples which were never extracted), and 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).

In order to select a robust filtering procedure, we computed a score which reflects the stability to data perturbation for the signal to noise and for the t-test, and chose that procedure for which the value of the score was higher. The score was defined as a variant R' of the robustness index R introduced in Stolovitzky (2003): In each experiment we selected the top 25% of genes using criteria described above. R' equals the ratio of the intersection of the genes selected which occur at least 90% of the time in the 500 experiments to the union of the genes selected by these experiments.

For an "ideal" robust 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 (where n << number of genes in the dataset) one expects R' to be small. To compute R' in this situation, we simulated the gene selection process assuming it was random. We found that R' is close to 0 whenever n < 500. R' reflects the stability of the filtering selection criterion to noise and to sample composition perturbations.

In our case, the robustness index associated with the signal to noise correlation was R'=0.27, while the robustness index of the t-test was R'=0.23. Based on these results, we chose the signal-to-noise correlation as a filtering criterion.

A pool of 73 genes passed the top 25% signal to noise filtering criterion in at least 90% of the perturbed datasets. We found that the selected pool contains only 51 of the 100 top genes selected by Shipp et al. (2002), 32 of the 100 genes selected by the method in Genes@Work (Stolovitzky, 2005), 28 genes selected based on the t-statistics (Stolovitzky, 2005), and only 9 p53 regulated genes. In addition, we found that in fact, only 25 genes from the Shipp et al. list remained in the list of top 100 in at least 90% of the perturbed training datasets.

The 95% CI of the absolute Pearson correlation coefficient among the pairs of the selected 73 genes was (0.33, 0,35) and only 19 pairs of genes had a correlation above 0.85 across all the samples. However, patterns extracted with Genes@Work revealed that several subsets of genes, e.g., G18, TXNIP, RPL13, NME1, TRIB2 have a low variation (δ < 0.1), being up-regulated on significant subgroups of about 30% DLBCL cases and down-regulated on subgroups of 50% FL cases (p-value 0.001). In fact, we were able to detect about 500 subsets of genes which collectively showed a variation < 0.1 on various subsets of 30-50% FL cases, and about 800 subsets of genes having a similar property on large subsets of DLBCL cases. The information provided by the groups of co-regulated genes is yet to be explored in a future study.

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.

Our support set was the collection of 30 (pairwise low-correlated) genes that occurred in the definition of the selected 57 patterns (see Table 1). Only 19 of these 30 genes were selected by Shipp et al. (2002) Out of the 11 genes in our set which were not included in Shipp et al. study, 10 are up-regulated in FLs and only one (STAT1) was 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., http://www.infobiogen.fr/services/chromcancer/); the remaining 4 genes are mostly involved in cell metabolism or transport.

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

The 19 genes in common with Shipp et al. (2002) 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.

Pattern data was defined with respect to the 57 selected patterns and can be visualized in Figure 3. 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.

Table 2. Examples of characteristic patterns for DLBCL and FL

Pattern	Gene symbol								Prevalence (%)			
									Training set		Test set	
	TXNIP	CDH11	GPR18	ELF2	MCM7	STAT1	PSMA6	CLTA	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

Meta-classifier construction and validation

We trained 6 individual classifiers (see Figure 1) 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 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.

Table 3 Performance of classifiers on training and test data.

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
	LR	0.07	97.37	84.62	5.88	94.12	69.23	12.77
Trained on pattern data	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**

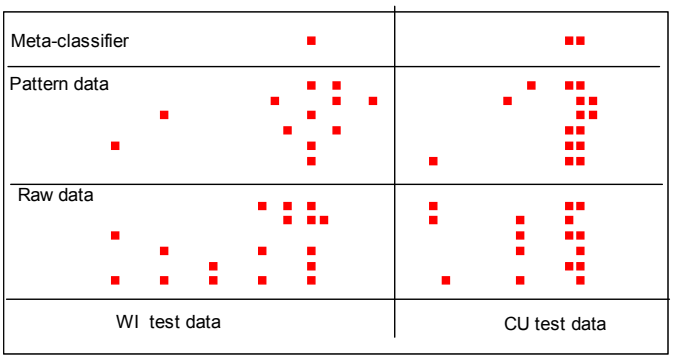

Figure 4. Error distribution of the meta-classifier and of the individual classifiers trained on raw and pattern data.

We further perturbed the raw training data with experimental noise, generated the corresponding pattern data, retrained 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 are presented in Figure 5. The fact that the meta-classifier predictions did not change is a confirmation of its robustness, in particular of its stability to experimental noise.

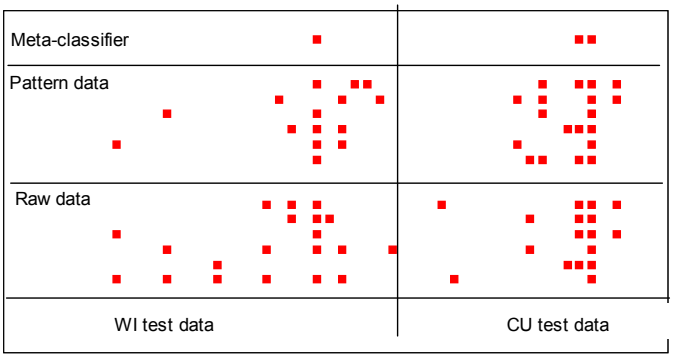

Figure 5. Error distribution of the meta-classifier and of the individual classifiers trained on perturbed data.

Biology-based gene selection: Role of p53 regulated genes

In a preliminary step, we identified 215 genes involved in biologically relevant pathways for p53 (Robins et al., in press) which were in both the WI and CU datasets. However, about 32% of these genes were not consistently regulated in DLBCL vs. FL cases in both WI and CU datasets, and only 90 genes were able to differentiate between DLBCL and FL with a p-value below 0.01 (see Table 4). 10 of these 90 genes were selected by Genes@Work (Lepre et al., 2004), 5 were selected by the t-test and by the signal to noise correlation criterion (Shipp et al., 2004) (Stolovitzky, 2005), and 4 genes (MCM7, BC2A1, CDNK2D and CCNG2) were selected in our support set of 30 genes. Four additional genes (LDHA, PGAM1, RPL13 and HSPCB) which were highly significant in differentiating between the lymphoma phenotypes in the WI data but were not meas-

ured accurately in the CU data and so were not included in our list.

Table 4 presents the list of top 90 p53 responsive genes which are significantly differentiating (p-value < 0.01) DLBCL vs. FL cases in both the WI and CU data. The genes which are up-regulated in the FL cases are marked with an asterisk (*).

Table 4. List of top 90 p53 responsive genes (p-value 0.01). The genes are listed in increasing order of their p-values (from $3x10^{-11}$ to 0.005).

Gene symbol			
CCNB1	EPRS	TOPBP1	CDK7
MCM7	GSK3B	PMAIP1	E2F3
BRCA1	COL6A1	ACAA2	MDM4
BCL2A1	HRAS	E2F5*	AMPD2
PPP2R4	SERPING1	POLA	RBBP4
EIF2S2	CCNA2	HMGB2	CCNG2*
COMT	CCT6A	PSMB5	HARS
IARS	MCM2	ACTA2	CASP6
MPI	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	

Our list of 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 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.

The core regulators of p53 which were identified in the WI and CU datasets are MDM2 and E2F1, and their expression levels are consistently up-regulated (p-value 0.10) and down-regulated (p-value 0.08) on the FL vs. DLBCL cases, respectively.

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. (Sander et al., 1993). If used as a biomarker, p53 alone can differentiate with a sensitivity of 70% and a specificity of 50%. From p53-dependent patterns from WI data, we identified that the best p53 responsive genes which were able to discriminate between the atypical p53 FL and DLBCL cases in our data were PLK1 and CDK2. Indeed, in 93% of the DLBCL cases with under-expressed p53 at least one of the genes PLK1 or CDK2 is up-regulated. In each FL case, at most one of the genes p53, PLK1 or CDK2 is up-regulated and in 58% of the FL cases none of these genes is over-expressed.

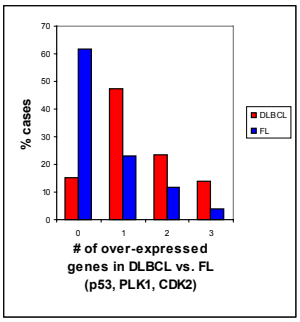

Figure 6. Histogram of % DLBCL (blue) and FL (red) cases having one up to three genes (p53, PLK1, CDK2) over-expressed.

Figure 6 depicts the histograms of the % cases DLBCL and FL for which 0, 1, 2, or 3 genes (p53, PLK1, CDK2) are over-expressed. Thus, p53, PLK1 and CDK2 might constitute a combinatorial biomarker for FL vs. DLBCL.

Among other significant individual phenotype biomarkers identified in our list we 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 p53 responsive genes may have even more predictive value. In Table 5 we present some examples of combinations of four p53 biomarkers which are characteristic for large subgroups of DLBCLs and FLs.

Table 5. Examples of high prevalence p53 patterns.

Pattern	Gene symbol				Prevalence (%)			
					Training set		Test set	
	MCM7	BCL2A1	JUN	ZNF184	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

Figure 7 presents the pattern data constructed on the WI training data and its performance on the CU test data. We found that MCM7, CCNG2, BCL2A1 and HLAE occur with high frequency (above 25%) in the definition of the patterns used in pattern data; in particular, MCM7 occurred in the definition of more than 90% of the patterns.

A weighted voting classifier trained on the patterns for the 90 p53 genes (see Table 4) for the WI data made the same 2 false positives on the CU test data as our meta-classifier, and one additional false negative (DLC14). We consider this is an interesting result, given that the gene selection in this case was imposed on the data on the basis of biological expectation that p53 is relevant.

In a forthcoming study we intend to explore these issues in greater detail.

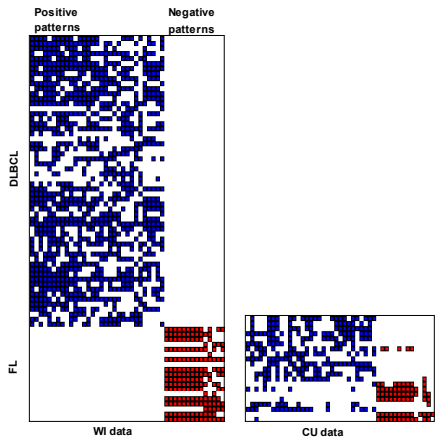

Figure 7. Visualization of p53-pattern data.

Meta-data analysis: extracting information from multiple support sets

In this section we present a method to integrate the predictions of different support sets. In general, such an analysis can be done using any classifier or meta-classifier.

We consider five different support sets: S1 is the support set selected by Shipp et al. (2002) which consists of 50 top correlated genes for each of the phenotypes DLBCL and FL. S2 and S3 are the support sets of Stolovitzky (2005) with 100 genes selected via a pattern-based method, and another 100 genes selected by t-test. S4 is the support set of our 30 genes discussed previously, and S5 is the support set of 90 p53 responsive genes. We use the weighted voting system as a predictor.

If we use WI as our training dataset, we find that the errors made by the weighted voting system on the 5 support sets are highly correlated and even after a principal component analysis, the results do not improve. Therefore, to illustrate the method, we use instead the meta-data obtained by combining the WI and CU datasets.

Table 6 presents the errors made by the weighted voting system on the five different support sets.

Table 6. Meta-classifier on multiple support sets.

Missclassified samples	S1	S2	S3	S4	S5	Meta-classifier
DLBCL 15	1		1		1	
DLBCL 21			1		1	
DLBCL 26	1	1	1		1	1
DLBCL 27			1			
DLBCL 29	**1**	**1**	**1**	**1**	**1**	**1**
DLBCL 35	1				1	
DLBCL 36		1				
DLBCL 39	1	1		1	1	1
DLBCL 40					1	
DLBCL 46				1		
DLBCL 52				1		
DLBCL 54					1	
DLBCL 56			1		1	
DLCL 7						
DLCL 13			1		1	
DLCL 14		1	1		1	
FL-DM						
FL-GL	**1**	**1**	**1**	**1**	**1**	**1**
Error rate	6	6	9	5	12	4
Weights	0.26	0.30	0.24	0.30	0.00	

As Table 6 shows a weighted combination of predictions from different support sets reduces the prediction error.

This example is meant to illustrate the general method for combining noisy results from different mathematical and statistical techniques and data from different laboratories into a better "meta-"predictor.

4 SUMMARY AND CONCLUSIONS

In this study we proposed a pattern-based meta-classification method for cancer detection from gene expression data and showed how it can differentiate between follicular lymphoma and diffuse large B-cell lymphoma on microarray data produced by two laboratories.

The method involves the selection of a robust subset of genes with low sensitivity to data perturbations produced by experimental noise or by altering the sample composition. The selected subset of genes is used to create predictions from several individual classifiers. The final phenotype prediction is obtained by integrating the individual classifications into a robust meta-classifier. We noticed that because the errors produced by the individual classifiers are uncorellated, the overall performance of the meta-classifier is superior to each individual predictor.

A novel approach used in our study was to train individual classifiers on pattern data, i.e., on a representation of the raw data in which significant patterns characteristic for the phenotype are viewed as synthetic variables. We showed that this approach lead to the increase of the performance of the individual classifiers.

Special attention was given to the role of p53 responsive genes in differentiating between follicular and diffuse large cell lymphomas. Although it is known that p53, PLK1 and CDK2 are each over-expressed in DLBCL or poor prognosis FL cases, we found that a decision based on combination of the expression levels of these three is a much more accurate predictor of phenotype. As Figure 6 shows, if none of them is elevated, the phenotype is FL 80% of the time, whereas if at least one of them is up-regulated, the phenotype is DLBCL 70% of the time.

ACKNOWLEDGEMENTS

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.

REFERENCES

Choudhuri S. (2004) Microarrays in biology and medicine. *J Biochem Mol Toxicol.* **18**(4),171-9.

Lyons P. (2003) Advances in spotted microarray resources for expression profiling. *Brief Funct Genomic Proteomic*.**2**(1), 21-30. Review.

Califano A, Stolovitzky G, Tu . (2000) Analysis of gene expression microarrays for phenotype classification. *Proc Int Conf Intell Syst Mol Biol*., **8**, 75-85.

Armstrong NJ, van de Wiel MA. (2004) Microarray data analysis: From hypotheses to conclusions using gene expression data. *Cell Oncol*. **26**(5-6), 279-90.

Slonim DK. (2002) From patterns to pathways: gene expression data analysis comes of age. *Nat Genet*. 32, 502-8. Review.

Bhanot G, Alexe G, Venkataraghavan B, Levine AJ. A robust meta-classification strategy for cancer detection from mass spectrometry data. *Submitted*.

Merz C. (1998) Classification and Regression by Combining Models. Disertation. UCI.

Prodromidis A L. Stolfo Salvatore J.A (1999) Comparative Evaluation of Meta-Learning Strategies over Large and Distributed Data Sets. Sixteenth International Conference on Machine Learning (ICML-99) 18-27, Bled Slovenia, June 1999.

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Med.; **8**(1), 68-74.

Stolovitzky G. (2005) *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 (preprint: http://www.wkap.nl/prod/a/Stolovitzky.pdf).

Lepre J, Rice JJ, Tu Y, Stolovitzky G. (2004) Genes@Work: an efficient algorithm for pattern discovery and multivariate feature selection in gene expression data. *Bioinformatics*;**20**(7), 1033-44.

Winter JN, Gascoyne RD, Van Besien K. (2004) Low-grade lymphoma. *Hematology* (Am Soc Hematol Educ Program), 203-20.

Moller MB, Nielsen O, Pedersen NT. (2002) Frequent alteration of MDM2 and p53 in the molecular progression of recurring non-Hodgkin's lymphoma. *Histopathology*, **41**(4):322-30

Pinyol M, Cobo F, Bea S, Jares P, Nayach I, Fernandez PL, Montserrat E, Cardesa A, Campo E. (1998) p16(INK4a) gene inactivation by deletions, mutations, and hypermethylation is associated with transformed and aggressive variants of non-Hodgkin's lymphomas. *Blood*, **91**(8), 2977-84.

Elenitoba-Johnson KS, Jenson SD, Abbott RT, Palais RA, Bohling SD, Lin Z, Tripp S, Shami PJ, Wang LY, Coupland RW, Buckstein R, Perez-Ordonez B, Perkins SL, Dube ID, Lim MS.(2003) Involvement of multiple signaling pathways in follicular lymphoma transformation: p38-mitogen-activated protein kinase as a target for therapy.*Proc Natl Acad Sci U S A*., **100**(12), 7259-64.

Lossos IS, Alizadeh AA, Diehn M, Warnke R, Thorstenson Y, Oefner PJ, Brown PO, Botstein D, Levy R. (2002) 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 U S A*. **99**(13), 8886-91.

Lossos IS, Jones CD, Zehnder JL, Levy R. (2001) A polymorphism in the BCL-6 gene is associated with follicle center lymphoma. *Leuk Lymphoma*. **42**(6), 1343-50.

McShane L., Radmacher RD, Freidlin B, Yu R, Li MC, Simon R.(2002) Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* **18**, 1462-69.

Alexe G, Alexe S, Hammer PL, Vizvari B. (2003) Pattern-based feature selection in genomics and proteomics. Rutgers University, RUTCOR Research Report RRR, **7**, 1-24.

Tu Y, Stolovitzky G, Klein U. (2002) Quantitative noise analysis for gene expression microarray experiments. *Proc Natl Acad Sci U S A*. **99**(22), 14031-36.

Stolovitzky G.(2003) Gene selection in microarray data: the elephant, the blind men and our algorithms. *Curr Opin Struct Biol*. **13**(3), 370-6. Review.

Alexe G, Hammer PL. (2005) Spanned patterns in Logical Analysis of Data. *Discr. Appl. Math*. in press.

Robins H, Alexe G, Harris S, Levine AJ. (2005) The first twenty-five years of p53 research. *Cell*, in press.

Finlay CA, Hinds PW, Levine AJ. (1989) The p53 proto-oncogene can act as a suppressor of transformation. *Cell*, **57**(7), 1083-93.

Monti S, Tamayo P, Mesirov J, Golub T. (2003) Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning J.* **52**(1-2), 91-118.

Matolcsy A, Warnke RA, Knowles DM. (1997) Somatic mutations of the translocated bcl-2 gene are associated with morphologic transformation of follicular lymphoma to diffuse large-cell lymphoma. *Ann Oncol*.**8** Suppl 2, 119-22.

McDonnell TJ, Deane N, Platt FM, Nunez G, Jaeger U, McKearn JP, Korsmeyer SJ (1989). bcl-2-immunoglobulin transgenic mice demonstrate extended B cell survival and follicular lymphoproliferation. *Cell*, **57**(1), 79-88.

Sander CA, Yano T, Clark HM, Harris C, Longo DL, Jaffe ES, Raffeld M. (1993) p53 mutation is associated with progression in follicular lymphomas.*Blood*. **82**(7), 1994-2004.

Lo Coco F, Gaidano G, Louie DC, Offit K, Chaganti RS, Dalla-Favera R. (1993) p53 mutations are associated with histologic transformation of follicular lymphoma. *Blood*, **82**(8), 2289-95.

Moller MB, Nielsen O, Pedersen NT. (1999) Oncoprotein MDM2 overexpression is associated with poor prognosis in distinct non-Hodgkin's lymphoma entities. *Mod Pathol*. **12**(11), 1010-6.