

# Differential Friendly Neighbors Algorithm for Differential Relationships Based Gene Selection and Classification using Microarray Data

R. Krishna Murthy Karuturi, Silvia Wong, Wing-Kin Sung and Lance D. Miller

**Abstract**—Identifying biologically relevant genes in a given tumor classification problem is as important as accurately classifying the samples. Differential expression may not always achieve this task. Differential friendly neighbors (DiffFNs) algorithm is proposed to select the classification relevant and biologically interesting genes that can discriminate two classes of tumors. DiffFNs achieves it by selecting genes based on their differential relationships from one class to the other.

DiffFNs has been applied to select genes that can discriminate patients based on their tumor p53 status and Grade status. The results show that DiffFNs identify the p53 activity associated genes which could be submerged in the differentially expressed genes if we used differential expression analysis alone. This result proves the effectiveness of DiffFNs approach in selecting genes that are biologically relevant to the sample classification.

**Keywords:** Gene selection, Microarray classification, Differential Friendly Neighbors, Survival analysis.

## I. INTRODUCTION

Several methods [8][9][20][1][25][17] have been proposed to find the molecular signature (a cassette of sample discriminating genes) and classify samples using microarray [22][3] genome-wide gene expression profiling. The signature gene selection is mostly carried out by choosing the genes which are differentially expressed among different classes i.e. the genes whose mean expression levels significantly changed from one class to the other class of samples are selected. The differentially expressed genes may be discovered by any of mean expression shift detecting methods such as t-test [11], SAM [24]. In practice, the top few differentially expressed genes are chosen as relevant features and are used for classifier design. One of the several classifiers such as KNN [5], SVM [4], NN [10], PAM [21] is used for classification

None of the above approaches address the problem of changing relationships among genes from one group of tumors to another group of tumors. The change of relationships among genes are important to identify the pathways that are activated or inactivated in a tumor type relative to the other

R. Krishna Murthy Karuturi is with Genome Institute of Singapore, Singapore (Phone: +65-64788040, Fax: +65-64789058, E-mail: karuturikm@gis.a-star.edu.sg)

Silvia Wong was with National University of Singapore, Singapore at the time of this work and she is currently with Silkroad Technologies, Singapore (Phone: +65-90680171, E-mail: silvia.wong@pendulab.com)

Wing-Kin Sung is with Genome Institute of Singapore and National University of Singapore, Singapore (Phone: +65-64788039, Fax: +65-64789058, E-mail: sungk@gis.a-star.edu.sg)

Lance D. Miller is with Genome Institute of Singapore, Singapore (Phone: +65-64788100, Fax: +65-64789060, E-mail: millerl@gis.a-star.edu.sg)

tumor type. This way of gene selection and tumor (or sample) classification is biologically important.

In this paper, we propose a new methodology called *differential friendly neighbors (DiffFNs)* to choose genes that show different relationships with the other genes among different tumors and propose a classifier which is based on the gene-gene relationships. DiffFNs is an extension of FNs [12] for supervised gene selection problem. The DiffFNs approach defines the relationship of two genes in a tumor type as the fraction of samples have the same direction of expression from their respective mean values. It finds the pairs of genes which lost or gained relationship from one type of tumor to another type of tumor. The genes which gained or lost most relationships are considered to be the most important genes for discriminating the tumors and signify certain pathways or activity. This approach can recognise genes whose mean expression shift, from one tumor to another tumor, is not statistically significant and use them in biologically meaningful way to discriminate the tumors.

Similar approaches have been adopted recently by Kostka & Spang [13] and Prieto et al [19]. Kostka & Spang discount for genes showing change of variability and Prieto et al identify genes showing change in variance. Both use heuristic stochastic optimization algorithms to identify such genes. Whereas, DiffFNs does not explicitly make any such assumptions about variance. Instead, it defines changed neighborhoods (or relationships) for each gene and arrives at the disturbed neighborhoods centered around genes called *base genes* which renders better interpretation of the results. The disturbed neighborhoods may contain genes with or without changed variance of expression or with changed co-expression. Apart from providing disturbed neighborhoods, DiffFNs also contain classification scheme based on these neighborhoods. Different gene expression centering schemes can give rise to different interpretation of the algorithm: (1) overall mean centering, a gene expression is adjusted such that its mean across all samples is zero but not the group mean, presented in this paper; and, (2) group mean centering, same as earlier but by making mean expression of a gene to be zero in each group individually. Unlike Kostka & Spang and Prieto et al, DiffFNs operate on binarized gene vectors and use dot product for similarity.

The rest of the paper is organized as follows. Section II formulates the *differential friendly neighbors (DiffFNs)* algorithm; presents the gene selection and classification methodologies based on it. Section III applies the DiffFNs approach

to classify grades and p53 status of breast cancers and evaluates the performance of our method. Section IV presents conclusions and future directions.

## II. DIFFERENTIAL FRIENDLY NEIGHBORS (DIFFFNs)

Consider  $A$  samples of class 1, denoted by  $S^1 = \{S^1_1, S^1_2, S^1_3, \dots, S^1_A\}$ , and  $B$  samples of class 2, denoted by  $S^2 = \{S^2_1, S^2_2, S^2_3, \dots, S^2_B\}$ . Their gene expression profiles over  $N$  genes can be represented by two matrices (datasets)  $D_1$  and  $D_2$  as

$$D_1 = (X_{ij})_{N \times A} \quad \text{and} \quad D_2 = (Y_{ij})_{N \times B}$$

where  $X_{ij}$  is expression of gene  $g_i$  in sample  $s_j \in S^1$  and  $Y_{ij}$  is expression of gene  $g_i$  in sample  $s_j \in S^2$ .

Different sample sizes in the two groups (i.e.  $A \neq B$ ) may bias the average expression level of a gene, across all tumors (samples), to the larger group. To get an unbiased estimation, we give equal weightage to both classes by taking mean of the means of the two classes. Formally, let  $M^1 = [M^1_1, M^1_2, \dots, M^1_N]^T$  and  $M^2 = [M^2_1, M^2_2, \dots, M^2_N]^T$  be the vectors of the means of expression of genes in datasets  $D_1$  and  $D_2$  respectively, where  $M^1_i$  and  $M^2_i$  are the mean expression values of the gene  $g_i$  in datasets  $D_1$  and  $D_2$  respectively and they are given by

$$M^1_i = \frac{1}{A} \sum_{j=1}^A X_{ij} \quad \text{and} \quad M^2_i = \frac{1}{B} \sum_{j=1}^B Y_{ij}.$$

then the overall mean

$$M = [M_1, M_2, \dots, M_N]^T = \frac{1}{2}(M^1 + M^2)$$

Now let us derive two new matrices  $\tilde{D}_1$  and  $\tilde{D}_2$  as

$$\tilde{D}_1 = (\tilde{X}_{ij})_{N \times A} \quad \text{and} \quad \tilde{D}_2 = (\tilde{Y}_{ij})_{N \times B}$$

where  $\tilde{X}_{ij} = U(X_{ij} - M_i)$ ,  $\tilde{Y}_{ij} = U(Y_{ij} - M_i)$ , and

$$U(z) = \begin{cases} 1 & \text{if } z \geq 0; \\ -1 & \text{if } z < 0. \end{cases}$$

Let the similarity between genes  $g_i$  and  $g_j$  in dataset  $D$  ( $\in \{\tilde{D}_1, \tilde{D}_2\}$ ), with  $Q \in \{A, B\}$  columns, be denoted by  $S(i, j|D)$  and defined as

$$S(i, j|D) = \frac{1}{Q} \sum_{k=1}^Q U(\tilde{Z}_{ik} \tilde{Z}_{jk})$$

In other words, if  $g_i$  and  $g_j$  are either induced together or repressed together from  $M_i$  and  $M_j$  respectively for a fraction  $\alpha$  of samples then  $S(i, j|D) = (2\alpha - 1) \in [-1, +1]$  for  $\alpha \in \{0, \frac{1}{Q}, \frac{2}{Q}, \dots, 1\}$ , which is a strictly increasing function of  $\alpha$ . This is a measure of degree of co-regulation of genes  $g_i$  and  $g_j$  from their respective mean expression values, in  $D$ . If  $g_i$  and  $g_j$  are positively (negatively) co-regulated then  $S(i, j|D)$  will be positive (negative). Further, if  $g_i$  and  $g_j$  are not co-regulated and their expressions follow

independent normal distributions, then the probability mass function  $f(\beta = S(i, j|D)) = \binom{Q}{\frac{Q(\beta+1)}{2}} \left(\frac{1}{2}\right)^Q$ .

Let  $T_1, T_2 \in [0, 1]$  be two parameters to be chosen by the user and they satisfy  $T_1 > T_2$ . Based on the above observation, we consider the expressions of two genes  $g_i$  and  $g_j$  to be independent or unrelated if  $-T_2 \leq S(i, j|D) \leq T_2$ . Otherwise, the gene pair  $(g_i, g_j)$  is said to have positive (or negative) relationship i.e. they are positive (or negative) friendly neighbors (FNs) of each other if  $S(i, j|D) \geq T_1$  (or  $S(i, j|D) \leq -T_1$ ). When  $S(i, j|D)$  is in  $[T_2, T_1]$  or  $[-T_1, -T_2]$ , we consider the relationship between  $g_i$  and  $g_j$  to be unclear. The following formulae give p-values,  $p^+$  ( $p^-$ ) of observing positive (negative) relationship between genes  $g_i$  and  $g_j$  when  $g_i$  and  $g_j$  are drawn independently from the respective normal distributions and  $S(i, j|D) = r$  is

$$p^+(g_i, g_j) = Pr(\beta \geq r) = \sum_{k=\frac{Q(r+1)}{2}}^Q \binom{Q}{k} \frac{1}{2^Q}$$

$$p^-(g_i, g_j) = Pr(\beta \leq r) = \sum_{k=0}^{\frac{Q(r+1)}{2}} \binom{Q}{k} \frac{1}{2^Q}$$

Let  $\beta_1 = S(i, j|D_1)$  and  $\beta_2 = S(i, j|D_2)$ . Then, table I summarizes the different changes of relationships between genes  $g_i$  and  $g_j$  from  $D_1$  to  $D_2$ .

TABLE I  
DEFINITION OF CHANGES OF RELATIONSHIPS USING THRESHOLDS ( $T_1$  AND  $T_2$ )

	$\beta_2 \geq T_1$	$\beta_2 \leq -T_1$	$\beta_2 \in [-T_2, T_2]$
$\beta_1 \geq T_1$	NC	P2N	$P_l$
$\beta_1 \leq -T_1$	N2P	NC	$N_l$
$\beta_1 \in [-T_2, T_2]$	$P_g$	$N_g$	NC

where

$P2N$  = Positive relationship of  $(g_i, g_j)$  changed to Negative

$N2P$  = Negative relationship of  $(g_i, g_j)$  changed to Positive

$P_g$  = Positive relationship of  $(g_i, g_j)$  is gained

$P_l$  = Positive relationship of  $(g_i, g_j)$  is lost

$N_g$  = Negative relationship of  $(g_i, g_j)$  is gained

$N_l$  = Negative relationship of  $(g_i, g_j)$  is lost

NC = No change in relationship of  $(g_i, g_j)$

Other cases when  $\beta_1, \beta_2$  are in intervals  $[T_2, T_1]$  and  $[-T_1, -T_2]$  are not evaluated since we would like to leave them as doubtful to decide whether the relationship exists or not. In practice, P2N and N2P are rare events and we will ignore them from now onwards.

## A. Gene Selection

Based on our theory, the genes that are relevant for discriminating samples into two given classes are the ones which gained or lost many friendly neighbors (FNs). We expect a gene involved in a pathway which has different activity in different tumor types is expected to loose or gain more FNs than the gene involved in other pathways. Hence, relevance of a gene for the given classification is a monotonic function of the number of FNs (positive or negative) gained or lost. In practice, almost all genes which gained many FNs usually have insignificantly fewer FNs lost i.e. we can take it granted that an important gene usually contributes to either gain or loss but not for both. In this analysis, we choose two types of genes: (1) *base genes* which either lost or gained many FNs; and, (2) *neighborhood genes* which are FNs to at least one of the base genes in one of the classes. The selection of base genes automatically defines the neighborhood genes. The score of a base gene is the maximum of the number of FNs gained and lost. If a base gene gained (lost) more FNs than lost (gained), we call such a gene as gain (loss) base gene. The base genes may be chosen in two ways using bootstrap technique [6]: (1) median gain or loss score of  $K$  bootstrap experiments; and, (2) number of FNs gained or lost consistently i.e. for more than 50% of the  $K$  bootstrap runs. Precisely, in each bootstrap run, randomly choose equal number of samples from each class and use the selected samples to compute, for each gene, the set of genes lost or gained. Let  $K$  be sufficiently large. For method 1, assign a score as median of scores in  $K$  runs. For method 2, retain all those pairs of genes whose relationship is observed to have occurred and be consistent in at least  $\lceil \frac{K}{2} \rceil$  runs.

## B. Sample Classification

We developed two classifiers, described below, based on the above described gene selection procedures. They are: (1) majority induction-repression classifier which is based on the first method of gene selection; and, (2) neighborhood based classifier which is based on the second method of gene selection.

1) *Majority induction-repression classifier*:: The genes chosen for this classifier is the highest (gain or loss) scoring genes in method 1 of gene selection. Any gene  $g_i$  chosen from gain (loss) analysis from  $D_1$  to  $D_2$  are assumed to exhibit expression variation in one direction from the mean expression ( $M_i$ ) consistently in  $D_2$  ( $D_1$ ) than in  $D_1$  ( $D_2$ ). We assume this consistency in  $D_2$  ( $D_1$ ) to be more than the required threshold ( $Th_1 \in [0, 1]$ ). The inconsistency of the expression of the same gene  $g_i$  in  $D_1$  ( $D_2$ ) should be less than  $Th_2 \in [-1, 1]$  and  $Th_1 \geq Th_2$  holds. This observation paves the way for our simple *majority induction-repression classifier* for classification of the samples. Note that, this assumption may not always hold since the FNs are defined based on the consistency of induction and repression across genes. This means, this classifier is valid to apply only when this assumption holds.

TABLE II  
PROCEDURE TO REDUCE THE BASE GENE SET  $\mathcal{B}$  WITH REDUNDANT NEIGHBORHOODS TO A BASE GENE SET  $\mathcal{B}^R$  WITH NON-REDUNDANT NEIGHBORHOODS.

---

INPUT: Ranked list of Base Genes  $\mathcal{B} = \{b_1, b_2, \dots, b_{nb}\}$   
and Neighborhoods  $Nbr = \{N^1, N^2, \dots, N^{nb}\}$   
OUTPUT: Reduced Base Gene List  $\mathcal{B}^R$

---

ALGM: ReduceBaseGenes( $\mathcal{B}, Nbr$ )  
 $\mathcal{B}^R = \{b_1\}$   
 foreach p (2:nb)  
   If  $N^p$  is not covered by any of the neighborhoods  
     of the base genes in  $\mathcal{B}^R$   
   Then  $\mathcal{B}^R \leftarrow \mathcal{B}^R \cup \{b_p\}$   
 end  
 Return( $\mathcal{B}^R$ )

---

Let  $V^g = [V_1^g, V_2^g, \dots, V_{v_g}^g]$  and  $V^l = [V_1^l, V_2^l, \dots, V_{v_l}^l]$  be profiles of gain and loss base genes, where

$$V_i^g = U\left(\sum_{j=1}^B \tilde{Y}_{ij}\right) \quad \text{and} \quad V_i^l = U\left(\sum_{j=1}^A \tilde{X}_{ij}\right)$$

Then a sample  $S = [S_1, S_2, \dots, S_N]^T$  whose transformation with respect to the mean vector  $M$  is  $\tilde{S} = [\tilde{S}_1, \tilde{S}_2, \dots, \tilde{S}_N]^T$ , where  $\tilde{S}_i = U(S_i - M_i)$ . Then, it gets two scores, one for each class i.e. ( $S^g, S^l$ ). Where

$$S^g = \frac{1}{v_g} \tilde{S}^T V^g \quad \text{and} \quad S^l = \frac{1}{v_l} \tilde{S}^T V^l$$

The rule to find the class ( $C_s$ ) of the sample  $S$  is

$$C_s = \begin{cases} 1 & \text{if } S^l \geq Th_1 \quad \text{and} \quad S^g \leq Th_2 \\ 2 & \text{if } S^g \geq Th_1 \quad \text{and} \quad S^l \leq Th_2 \\ None & \text{Otherwise} \end{cases}$$

If the above assumption does not hold on any of the datasets then the clauses involving this dataset have to be removed from the above classification rule. This means, if the assumption does not hold on  $D_1$  ( $D_2$ ), the clauses involving  $S^l$  ( $S^g$ ) have to be deleted from the rule.

2) *Neighborhood based classifier*:: In this approach, base genes are chosen using the 2nd gene selection method. For each base gene  $g_i$ , it is associated with the respective neighborhood of genes given by the set  $N^i$ . Same neighborhood may be shared by different base genes. To eliminate this redundancy, we rank the base genes in the descending order of their strengths i.e.  $|N^i|$  of their neighborhoods. Starting from the second strong base gene, check whether the base gene's neighborhood is shared significantly by any of the base gene which is still in the list of base genes and its neighborhood strength is higher. If so, eliminate such a base gene from the list of base genes. The algorithm is formally described in Table II.

The neighborhood  $N^i$  is said to be covering the neighborhood  $N^j$  if

$$\frac{|N^i \cap N^j|}{|N^j|} \geq T_n \in [0, 1]$$

for a given threshold parameter  $T_n$ . After having reduced  $\mathcal{B}$  to  $\mathcal{B}^{\mathcal{R}}$ , we define two matrices  $G$  and  $L$ , for gain and loss respectively, to indicate the relationships between base genes and their neighborhood genes. Any element,  $G_{ij}$  or  $L_{ij}$ , from these two matrices take a value from the set  $\{-1, 0, +1\}$ . The values are defined as follows.

$$G_{ij} = \begin{cases} +1 & \text{if } g_i \text{ is a base gene in } \mathcal{B}^{\mathcal{R}} \text{ and} \\ & g_j \text{ is its neighborhood gene and} \\ & (g_i, g_j) \text{ is a } P_g \text{ from } D_1 \rightarrow D_2; \\ -1 & \text{if } g_i \text{ is a base gene in } \mathcal{B}^{\mathcal{R}} \text{ and} \\ & g_j \text{ is its neighborhood gene and} \\ & (g_i, g_j) \text{ is a } N_g \text{ from } D_1 \rightarrow D_2; \\ 0 & \text{Otherwise.} \end{cases}$$

$$L_{ij} = \begin{cases} +1 & \text{if } g_i \text{ is a base gene in } \mathcal{B}^{\mathcal{R}} \text{ and} \\ & g_j \text{ is its neighborhood gene and} \\ & (g_i, g_j) \text{ is a } P_l \text{ from } D_1 \rightarrow D_2; \\ -1 & \text{if } g_i \text{ is a base gene in } \mathcal{B}^{\mathcal{R}} \text{ and} \\ & g_j \text{ is its neighborhood gene and} \\ & (g_i, g_j) \text{ is a } N_l \text{ from } D_1 \rightarrow D_2; \\ 0 & \text{Otherwise.} \end{cases}$$

Now we define a binary operator denoted by,  $\underline{\Delta}$ , on two  $N \times N$  matrices  $T$  and  $U$  resulting in an  $N \times 1$  vector  $R(T, U) = T \underline{\Delta} U = [R_1, R_2, \dots, R_N]^T$ , where

$$R_i = \frac{\sum_{j=1}^N (U(A_{ij}B_{ij}) + 1)}{\sum_{j=1}^N (U(-A_{ij}B_{ij}) + 1)}$$

Performing  $G \underline{\Delta} (\tilde{S} \tilde{S}^T)$  and  $L \underline{\Delta} (\tilde{S} \tilde{S}^T)$  gives two vectors  $G^s$  and  $L^s$  respectively.  $G_i^s$  gives the ratio of the number of agreements and the number of disagreements in the base gene  $g_i$  and its neighborhood gene relationships between the sample and the neighborhood matrix  $G$ . Similar interpretation holds for  $L_i^s$ .

Now the sample has to be classified into either class 1 or class 2 or *None* based on the vectors  $G^s$  and  $L^s$ . We treat each  $G_i^s$  ( $L_i^s$ ) as a classifier, where gene  $g_i$  is a base gene in  $\mathcal{B}^{\mathcal{R}}$ .

Based on our gene selection criterion, we expect the following relationships to hold for  $i \in \{1, 2\}$

$$\frac{Pr(U(S_a S_b) = G_{ab} | C_s = i, G_{ab} \neq 0)}{Pr(U(S_a S_b) \neq G_{ab} | C_s = i, G_{ab} \neq 0)} \geq \frac{1 + T_{3-i}}{1 - T_{3-i}}$$

$$\frac{Pr(U(S_a S_b) = L_{ab} | C_s = i, L_{ab} \neq 0)}{Pr(U(S_a S_b) \neq L_{ab} | C_s = i, L_{ab} \neq 0)} \geq \frac{1 + T_i}{1 - T_i}$$

The above observations are the basis for the class assigned to the sample  $S$  by a given base gene,  $g_i$ . These can be translated to the following classification rules for a base gene  $g_i$  in  $G$  and  $g_j$  in  $L$  are

$$C_s^{ig} = \begin{cases} 2 & \text{if } G_i^s \geq \frac{1+T_1}{1-T_1} \\ 1 & \text{if } G_i^s \leq \frac{1+T_2}{1-T_2} \\ None & \text{Otherwise} \end{cases}$$

Similarly,  $L_i^s$  classifies the sample  $S$  according to the similar rule i.e.

$$C_s^{jl} = \begin{cases} 1 & \text{if } L_j^s \geq \frac{1+T_1}{1-T_1} \\ 2 & \text{if } L_j^s \leq \frac{1+T_2}{1-T_2} \\ None & \text{Otherwise} \end{cases}$$

Let  $S_C$  be defined as

$$S_C = \frac{|\{k/C_s^{kg}=1 \text{ or } C_s^{kl}=1, g_k \in \mathcal{B}^{\mathcal{R}}\}|}{|\mathcal{B}^{\mathcal{R}}|} - \frac{|\{k/C_s^{kg}=2 \text{ or } C_s^{kl}=2, g_k \in \mathcal{B}^{\mathcal{R}}\}|}{|\mathcal{B}^{\mathcal{R}}|}$$

The rule for the overall class  $C_s$  assigned to the sample  $S$  is

$$C_s = \begin{cases} 1 & \text{if } S_C \geq T_m \\ 2 & \text{if } S_C \leq -T_m \\ None & \text{Otherwise} \end{cases}$$

The above classification methodology does not depend on the distribution of the samples among the classes 1 and 2. This classifier does not assume that the gene to induce (or repress) consistently across all samples in at least one data set.

### III. EVALUATION & RESULTS

We applied the proposed methodology to the breast tumor affymetrix array data [16]. We used these procedures to find differentially behaving genes to discriminate histologic grade and p53 gene status.

#### A. Data Description & Problem Background

The data contains 257 affymetrix arrays, each array profiled the gene expression of one breast cancer patient. Each array was annotated for the p53 gene mutational status (p53+, p53-), histologic grade (Grade1, Grade2, Grade 3) and the survival information. Table III summarized the data distribution into various classes and combinations of classes.

Human breast tumors are diverse in their natural history and in their responsiveness to treatments. In general, different breast cancer patient will get different treatment depending on the histologic features of the breast cancer, such as its type and grade. Treatment decision will also be based on the phenotypic characteristic of the tumor, such as the presence of estrogen receptor (ER+) and HER-2/neu oncogene status. In this paper, we will be looking at two of the important relevant prognostic and treatment-guiding markers: (1) p53 status, an apoptotic gene acts as a break on cell proliferation whose loss of function may cause uncontrolled growth of abnormal cells leading to poor prognosis [15]; and, (2)

TABLE III  
DATA DISTRIBUTION INTO VARIOUS TUMOR STATUS CLASSES

Status	p53+	p53-	G1	G2	G3
p53+	59	0	3	24	31
p53-		198	66	105	23
G1			70	0	0
G2				129	0
G3					55

Grade, divided in to G1, G2 and G3 in the order of their severity according to The Scarf-Bloom-Richardson system [7].

### B. Methods

We take both direct and indirect evaluation to assess our methodology on the above dataset. The direct evaluation is misclassification rate. The indirect evidence is the classifier's ability to provide a classification which can correlate the redefined classification with the survival. The basis of this survival prediction is that pathologist's decision on the above classification may not always be reliable. For example, p53 status indicates only the DNA sequence level mutation status. The function of p53 may be inactivated because of several reasons that can be attributed to its up/down-stream genes in its pathway or to the post-translational modification. For p53 related assessment of treatment and survival, we need to assess whether the functionality of p53 in the context of breast cancer is active. Similar arguments hold for the grade status of the tumor. Hence we give more weightage to the survival separability than mere misclassification rate as in [16]. Moreover, we allowed some samples to be assigned no class since the status information is not really of binary in nature. We allowed our classifier to treat the border cases as unclassified.

To assess the survival separability of a certain classification of samples, we use Cox model Proportional Hazard (CoxPH) p-value [14] and likelihood ratio (LHR) test [18] between two classes. A Cox model provides an estimate of the treatment effect on survival after adjustment for other explanatory variables. It allows us to estimate the hazard (or risk) of death, or other event of interest, for individuals, given their prognostic variables. From a set of observed survival times (including censored times) in a sample of individuals, we can estimate the proportion of the population of such people who would survive a given length of time under the same circumstances. This method is called the product limit or Kaplan-Meier method.

### C. Results

We applied DiffFNs method both on Grade3-Grade1 pair (i.e.  $D_1 = \text{Grade3}$ ,  $D_2 = \text{Grade1}$ ) and on (p53+)-(p53-) pair (i.e.  $D_1 = \text{p53+}$ ,  $D_2 = \text{p53-}$ ). We developed classifiers independently using the genes with the relationships gained from G3 to G1 and p53+ to p53-. The reason for this choice is that in Grade1/p53- there is a control over cancer progression which may be lost in Grade3/p53+. These relationships signify the processes which are to be active to keep the cell differentiation and proliferation under control whose loss can lead to cancer.

a) *Majority Induction-Repression Classifier*:: We performed independent analyses of discriminating Grade3 from Grade1 and p53+ from p53-. We have chosen the base genes whose median gain of FN's from Grade3 to Grade1 is more than 100 which resulted in 1105 genes. Similarly, we have chosen base genes for p53+ to p53- case which resulted in 457 genes. The number of common genes among these two sets is 358. Since the overlap is so high, we built the classifiers based on these 358 genes. The classifier with the threshold parameters of  $Th_1 = Th_2 = 0$  was designed. Table IV shows both misclassification and independent survival analysis. The performance on Grade is not as good as in p53 case. The p-value of separation of survival times has improved by a factor of 100 and the likelihood ratio has improved by a factor of 2.5.

b) *Neighborhood classifier*:: We applied our methodology with thresholds  $T_1 = 0.6, T_2 = 0.3, T_n = 0.6$ , and  $T_m = 0.7$  to discriminate Grade3 from Grade1 and p53+ from p53-. The analysis carried out using these independent sets does not show good improvement of the predictability of the survival.

So, we selected those relationships which are gained from Grade3 to Grade1 as well as from p53+ to p53- and conducted the neighborhood reduction and classification analysis. We have chosen the base genes whose neighborhoods contain at least 30 genes. This resulted in 4 base genes with neighborhood strengths, in the decreasing order, 105,99,33 and 30. The complete results are shown in Table V. The results here are as surprisingly better as shown in the case of *Majority Induction-Repression Classifier* above except on grade.

Of the four neighborhoods selected above, the base genes with a lot of neighborhood genes are enriched in cell cycle associated genes. For example, the neighborhood of CRK7 contains 10 cell-cycle associated genes. This enrichment is significant at the p-value of  $1.69 \times 10^{-8}$ . The neighborhood of the base gene PPARBP is enriched in cancer associated genes, it is significant at the p-value of  $10^{-4}$ . We used cancer gene database (<http://caroll.vjf.cnrs.fr/cancergene/RETRI1.html>) for the analysis of this neighborhood. It does not have any cell cycle associated genes in its neighborhood. PPARBP itself is a breast cancer related gene which is known to get amplified in breast cancers and to regulate the p53 associated apoptosis. Of these 10 cancer related genes, four are p53 associated

TABLE IV

CLASSIFICATION AND SURVIVAL PREDICTION SUMMARY FOR *Majority Induction-Repression Classifier*.  $Th_1 = Th_2 = 0$ ,  $T_1 = 0.6$ ,  $T_2 = 0.3$ . THE NUMBERS SHOWN IN THE PARATHESSES ARE THE CLASSIFICATION OF GRADE2 INTO GRADE1 AND 3.

Group1	Grade3	p53+
Group2	Grade1	p53-
Misclassified in Group1	4 (32)	22
Misclassified in Group2	3 (97)	34
Unclassified in Group1	0	0
Unclassified in Group2	0	0
Original CoxPH p-value	$3 \times 10^{-4}$	$7 \times 10^{-3}$
New CoxPH p-value	$6 \times 10^{-5}$	$6 \times 10^{-5}$
Original LHR	15.2	6.63
New LHR	15.6	16.1
Original LHR P-value	$1 \times 10^{-4}$	0.01
New LHR P-value	$8 \times 10^{-5}$	$6 \times 10^{-5}$

genes whose p-value is  $3 \times 10^{-3}$  i.e. probability of getting at least 4 p53 related genes if we randomly picked up 10 cancer associated genes from this database of 2600 cancer genes.

We examined the misclassification rate of each of those neighborhoods and found that the PPARBP neighborhood gives the lowest misclassification rate consistently for all datasets i.e. Grade and p53. To examine its ability to predict the survival, we used PPARBP and its neighborhood alone as a classifier and the results are shown in Table VI. It shows that the results on p53 label prediction and survival are better than the ensemble of base gene classifiers while the performance on grade degraded. This could be because PPARBP and its neighborhood are specific to p53 than the grade.

#### IV. CONCLUSIONS & FUTURE WORK

We proposed a novel supervised gene selection and classification scheme called *Differential Friendly Neighbors* algorithm. Unlike conventional approaches which discover genes differentially expressed in different groups of samples, this algorithm discovers genes which exhibit differential relationships with the other genes among different groups of samples. The approach to classify samples is also different from the regular approach of gene-expression based classification. The approach used in DiffFNs is gain or loss

TABLE V

CLASSIFICATION AND SURVIVAL PREDICTION SUMMARY FOR *Neighborhood classifiers* DEVELOPED BASED ON THE COMMON NEIGHBORHOODS BETWEEN GRADE3 TO GRADE1 AND p53+ TO p53-.  $T_1 = 0.6$ ,  $T_2 = 0.3$ ,  $T_n = 0.6$ , and  $T_m = 0.7$ . THE NUMBERS SHOWN IN THE PARATHESSES ARE THE CLASSIFICATION OF GRADE2 INTO GRADE1 AND 3.

Group1	Grade3	p53+
Group2	Grade1	p53-
Misclassified in Group1	4 (14)	2
Misclassified in Group2	1 (50)	10
Unclassified in Group1	33	37
Unclassified in Group2	16	79
Original CoxPH p-value	0.0003	0.007
New CoxPH p-value	0.002	0.0002
Original LHR	15.2	6.6
New LHR	9.4	14.4
Original LHR P-value	$10^{-4}$	0.01
New LHR P-value	0.002	0.00015

of gene-gene relationships as attributes for classification. We applied our methodology to reclassify p53 status whose labels given by pathologists could be wrong and shown that the reclassification of these samples into p53- and p53+ did significantly better than the original labels assigned by the pathologists. DiffFNs method found a gene, PPARBP, that is involved in p53-dependent apoptosis to be the best base gene and the classifier built performed very well on p53 label prediction. The neighborhood of PPARBP involves four p53 associated genes and 6 other cancer associated genes whose significance would have been buried had we used traditional statistical tests like 2-group t-test.

The methodology needs to be improved in the sense of reducing base genes by automatically finding non-overlapping neighborhoods which could be more optimal than the one presented in this paper. The proposed method has to be further validated on other datasets. The method has to be examined to study the effect of various parameters on the results and the data properties like mislabeling also.

#### V. ACKNOWLEDGEMENTS

We thank Joshy George, Vega and Edison Liu for their valuable and timely suggestions during this work.

TABLE VI

CLASSIFICATION AND SURVIVAL PREDICTION SUMMARY FOR *Neighborhood classifier* DEVELOPED BASED ON PPARBP'S COMMON NEIGHBORHOOD BETWEEN GRADE3 TO GRADE1 AND P53+ TO P53-.  $T_m = 0.2$ . THE NUMBERS SHOWN IN THE PARATHESSES ARE THE CLASSIFICATION OF GRADE2 INTO GRADE1 AND 3.

Group1	Grade3	p53+
Group2	Grade1	p53-
Misclassified in Group1	2 (60)	0
Misclassified in Group2	8 (38)	60
Unclassified in Group1	4	3
Unclassified in Group2	13	48
Original CoxPH p-value	0.0003	0.007
New CoxPH p-value	0.006	$2 \times 10^{-4}$
Original LHR	15.2	6.6
New LHR	8	18.2
Original LHR P-value	0.0001	0.01
New LHR P-value	0.005	$2 \times 10^{-5}$

## REFERENCES

- [1] Alon, U. et al. (1999), Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays, *PNAS USA*, 96, 6745-6750.
- [2] Antanov, A.V. et al. (2004), Optimization models for cancer classification: extracting gene interaction information from microarray expression data, *Bioinformatics*, 20(5), 644-652.
- [3] Barrett, J.C. and Kawasaki, E.S. (2003), Microarrays: the use of Oligonucleotides and cDNA for the Analysis of Gene Expression, *Drug Discovery*, 8, 134-141.
- [4] Burges, C.J.C (1998), A Tutorial on Support Vector Machines for Pattern Recognition, *Knowledge Discovery and Data Mining*, 2(2).
- [5] Duda, R.O, Hart, P.E. and Stork, D.G. (2000), *Pattern Classification*, John Wiley & Sons, New York.
- [6] Efron, B. and Tibshirani, R.J. (1994), *An Introduction to the Bootstrap* Chapman & Hall, New York.
- [7] Elston, CW, (1987), Grading of invasive carcinoma of the breast, *Diagnostic Histopathology of Breast*, eds Page DL and Anderson TJ, Churchill Livingstone, 300-311.
- [8] Getz, G. et al. (2000), Coupled Two-way Clustering Analysis of Gene Microarray Data, *PNAS USA*, 97, 12079-12084.
- [9] Golub, T.R. et al. (1999), Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, 286, 531-537.
- [10] Haykin, S.S. (1999), *Neural Networks: A Comprehensive Foundation*, Prentice Hall, Upper Saddle River.
- [11] Kanji, G.K. (1999), *100 Statistical Tests*, SAGE Publications.
- [12] Karuturi R.K.M. and Vinsensius V.B. (2004), Friendly Neighbors Method for Unsupervised Determination of Gene Significance in Time-course Microarray Data, in the proc of *5th IEEE Symposium on Bioinformatics and Bioengineering*.
- [13] Kostka D. and Spang R. (2004), Finding disease specific alterations in the co-expression of genes, *Bioinformatics*, 20:i194 - i199.
- [14] Lawless, J.F. (1982), *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York.
- [15] Linjawi, A. et al. (2004), Prognostic Significance of p53, bcl-2, and Bax Expression in early breast cancer, *J Am Coll Surg*, 198(1), 83-90.
- [16] Miller, L.D. et al. (2004), An Expression Signature for p53 Status in Human Breast Cancer Predicts Mutation Status, Transcriptional Effects, and Patient Survival, *PNAS*, 102:13550-13555, 2005.
- [17] Neo, S.Y. et al. (2004), Identification of discriminators of hepatoma by gene expression profiling using minimal dataset approach, *Hepatology*, 39(4),944-953.
- [18] Parmar, M.K.B and Machin, D. (1995), *Survival Analysis: A Practical Approach*, John Wiley & Sons, New York.
- [19] Prieto C. et al. (2006), Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes, *Bioinformatics*, 22:1103-1110.
- [20] Ramaswamy, S. et al. (2001), Multiclass Cancer Diagnosis Using Tumor Gene Expression Signature, *PNAS USA*, 98, 15149-15154.
- [21] Tibshirani, R.O. et al. (2002), Diagnosis of Multiple Cancer Types by Shrunk Centroids of Gene Expression, *PNAS USA*, 99(10), 6657-6572.
- [22] Schena, M. et al. (1995), Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270, 467-470.
- [23] Shevade, S.K. and Keerthi, S.S. (2003), A Simple and Efficient Algorithm for Gene Selection Using Sparse Logistic Regression, *Bioinformatics*, 19, 2246-2253.
- [24] Tusher, V.G. et al. (2001), Significance Analysis of Microarrays Applied to the Ionizing Radiation Response, *PNAS USA*, 98, 5116-5121.
- [25] van't Veer, L.J. et al. (2002), Gene Expression Profiling Predicts Clinical Outcome Of Breast Cancer, *Nature*, 415, 530-536.