

Queensland University of Technology
Brisbane Australia

This may be the author's version of a work that was submitted/accepted for publication in the following source:

Peterson, Leif, Ozen, Mustafa, Erdem, Halime, Amini, Andrew, Gomez, Lori, Nelson, Colleen, & Ittmann, Michael
(2005)
Artificial neural network analysis of DNA microarray-based prostate cancer recurrence. In
2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2005-11-14 - 2005-11-15.

This file was downloaded from: https://eprints.qut.edu.au/9326/

© Copyright 2005 IEEE

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Notice: *Please note that this document may not be the Version of Record (i.e. published version) of the work. Author manuscript versions (as Submitted for peer review or as Accepted for publication after peer review) can be identified by an absence of publisher branding and/or typeset appearance. If there is any doubt, please refer to the published source.*

Artificial Neural Network Analysis of DNA Microarray-based Prostate Cancer Recurrence

Leif E. Peterson, Mustafa Ozen, Halime Erdem, Andrew Amini,
Lori Gomez, Colleen C. Nelson, and Michael Ittmann

***Abstract*— DNA microarray-based gene expression profiles have been established for a variety of adult cancers. This paper addresses application of an artificial neural network (ANN) with leave-one-out testsing and 8-fold cross-validation for analyzing DNA microarray data to identify genes predictive of recurrence after prostatectomy. Among 725 genes screened for ANN input, a 16-gene model resulted in 99-100% diagnostic sensitivity and specificity: *DGCR5, FLJ10618, RIS1, PRO1855, ABCB9, AK057203, GOLGA5, HARS, AK024152, HEP27, PPIA, SNRPF, SULT1A3, SECTM1, EIF4EBP1*, and *S71435*. Genes identified with ANN that are prognostic of prostate cancer recurrence may be either causal for prostate cancer or secondary to the disease. Nevertheless, the genes identified may be confirmed in the future to be markers of early detection and/or therapy.**

I. Introduction

Prostate cancer is the 2nd leading cause of cancer mortality among US males[1]. For locally confined cancer, curative treatment includes radical prostatectomy [2,3] and radiotherapy[4-7]. Following prostatectomy, it is expected that serum levels of prostate specific antigen (PSA) remain fixed below a level of 0.2 ng/ml. However, approximately 15-35% of patients experience PSA (biochemical) recurrence defined as two or more successive follow-up values of PSA>0.2 ng/ml more than 30 days post-surgery [8-11]. Early biochemical recurrence occurring before 1 year post-surgery is strongly associated with metastatic disease in approximately 75% of male patients[12]. However, clinical and pathological factors are imperfect predictors of biochemical recurrence. Thus, it is of interest to augment clinical information with prognostic information derived from molecular markers of aggressive disease.

Research for identifying biomarkers as a supplement to clinical information for prostate cancer has increased tremendously over the last decade. With regard to gene expression studies on recurrence, Lapointe et al identified a set of 23 genes whose expression levels were either positively or negatively associated with early recurrence in a set of 7 prostate cancers with early recurrence and 22 non-recurrent tumors[13]. Singh et al [14] studied gene expression of 12,600 genes in 21 prostate cancers (8 recurrent, 13 non-recurrent), and defined a set of 5 genes whose expression levels resulted in 90% accuracy of class prediction. Glinsky et al [15] performed additional xenograft-based analyses on the same data set and derived three five-gene models resulting in 95% classification accuracy. Lastly, Yu and colleagues [16] looked at a 70-gene leave-one-out classification model for 29 aggressive tumors and 37 non-aggressive tumors and obtained 86% classification accuracy.

There is growing interest in the application of ANNs in microarray analysis[17-21]. The majority of these papers conclude unique and superior classification results obtained with ANNs when compared with other classifiers. In this report, we applied an artificial neural network (ANN) to identify microarray-based genes with the greatest class-specific sensitivity for predicting PSA recurrence of prostate cancer. Two-color oligonucleotide microarrays were employed for generating patient gene expression profiles. The set of genes identified may ultimately be used to provide additional prognostic information in men with prostate cancer that will be useful for evaluation of prognosis and treatment planning.

II. Methods

A. Microarray Samples and Gene Ranking

We used a total of 32 arrays for prostate cancer cases undergoing radical prostatectomy. RNA was extracted from the peripheral zone in tissue with greater than 70% tumor pathology. Two-color 21k gene oligonucleotide arrays were hybridized and scanned for 15 cases which experienced post-surgical PSA recurrence <1 year ("poor outcome") and 17 cases which remained recurrent-free more than 5 y ("good outcome") post-surgery. From 21,329 genes, we identified 5,757 which had at least 95% informative spots (G and R signal-to-noise ratio>2) within each class. The non-parametric Mann-Whitney U test, which approximates the Gini diversity index [23], of the ChipST2C program (http://www.chipst2c.org) was used to identify 725 genes with significant rank differences (P <0.01) in

Correspondence: Leif E. Peterson, 1709 Dryden Street, Suite 1025, Houston, Texas 77030, USA.

Telephone: +001 (713) 798-5385, E-mail: peterson@bcm.tmc.edu.

Leif E. Peterson is with the Dept. of Medicine and Dept. of Molecular and Human Genetics, and Mustafa Ozen, Halime Erdem, Andrew Amini, Lori Gomez, and Michael Ittmann are with the Dept. of Pathology, Baylor College of Medicine, Houston, Texas 77030, USA.

Colleen C. Nelson is with the Prostate Centre at Vancouver General Hospital and Dept. of Surgery, Univ. of British Columbia, Vancouver, BC Canada V6T 1Z4.

0-7803-9387-2/05/$20.00 ©2005 IEEE

expression between poor and good outcome. Gene ranking resulted in a $32 \times 725(A \times G)$ **E** matrix of standardized expression, where $A(a = 1, 2, \ldots, 32)$ is the number of samples and $G(g = 1, 2, \ldots, 725)$ is the number of genes.

B. Dimensional Reduction

Dimensional reduction using k-means cluster analysis and principal component analysis (PCA) was performed with the ANNChip computer program (http://www.chipst2c.org). The objective for running k-means and PCA was to reduce the number of input nodes to be clamped to the ANN during training. In addition, individual gene expression values were mapped back to the k-means centers and PC scores by use of k-means scores and PC score coefficients. During k-means clustering in ANNChip, the optimum number of k clusters occurs when the ratio of the smallest between-cluster distance to total mean-square error (grand total of error between each gene and its mean vector) is the greatest. K-means cluster analysis resulted in a $32 \times 27(A \times K)$ **M** matrix of cluster centers, where $A(a = 1, 2, \ldots, 32)$ is the number of samples and $K(k = 1, 2, \ldots, 27)$ is the number of centers. Figure 1 shows a countour plot of the **M** matrix of 27 centers for the 32 arrays.

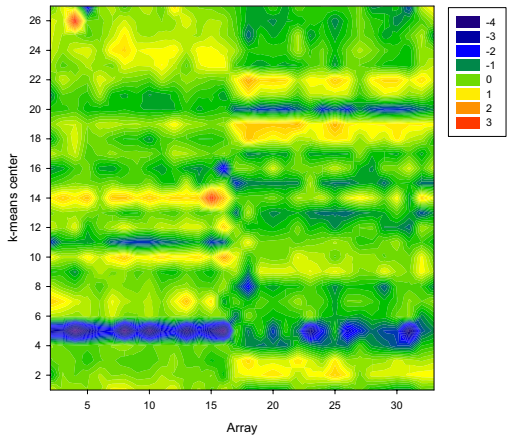


Fig. 1. **M** matrix of k-means centers for $k = 27$ clusters extracted from the expression values of 725 genes on 32 arrays.

Because the ANN is trained on samples with 27 input values representing the k=27 clusters, the gene expression profiles over 32 arrays needed to be mapped to the average expression within the 27 clusters by using "k-means scores."

The k-means score which maps gene g to center k is determined as

$$z_{gk} = \frac{\|\mathbf{x}_g - \mathbf{m}_k\| - \mu_k}{\sigma_k} \qquad k = 1, 2, \ldots, K, \qquad (1)$$

where $\mathbf{x}_g$ is the standardized expression vector for gene g, $\mathbf{m}_k$ is the mean vector for center k, $\|\mathbf{x}_g - \mathbf{m}_k\|$ is the Euclidean distance between expression for gene g and center k, and μ_k and σ_k are the average and standard deviation of distances $\|\mathbf{x}_g - \mathbf{m}_k\|$ between all genes and center k. This was repeated for each cluster center to yield a $725 \times 27(G \times K)$ **Z** matrix of k-means scores.

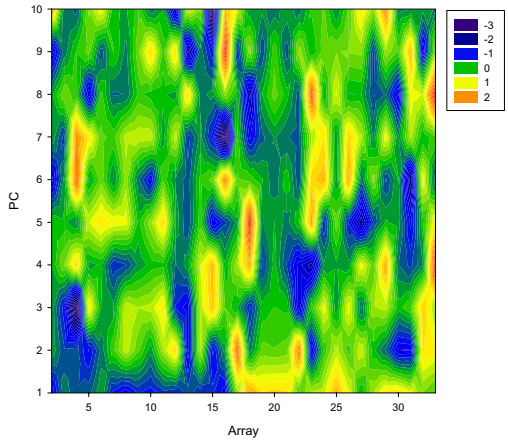


Fig. 2. **F** matrix of principal component scores for 10 PCs extracted from the expression values of 725 genes on 32 arrays. Total variance explained is 73.4%.

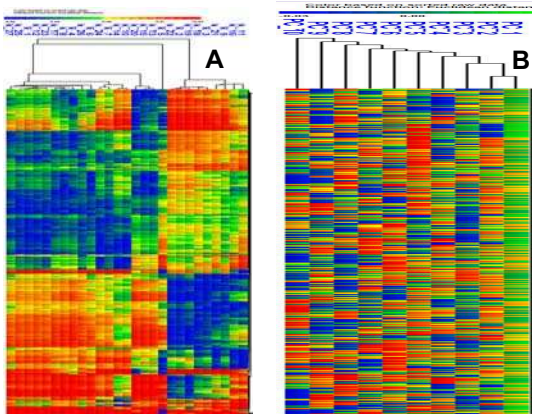


Fig. 3. (A) **Z** matrix of k-means scores and (B) **W** matrix of PC score coefficients for 725 genes, which map expression back to the k-means centers and PC scores.

During PCA dimensional reduction, the top 10 eigenvalues explaining 73.4% of total variation were extracted from the $725 \times 725(G \times G)$ correlation matrix **R**. PCA resulted in a $32 \times 10(A \times P)$ **F** matrix of PC scores, where $A(a = 1, 2, \ldots, 32)$ is the number of samples and $P(p = 1, 2, \ldots, 10)$ is the number of PCs. Figure 2 illustrates a countour plot of the **F** matrix of 10 PCs for the 32 arrays. Figure 3 shows the $725 \times 27(G \times K)$ **Z** matrix of k-means scores and the $725 \times 10(G \times P)$ **W** matrix of PC score coefficients, which map expression back to the k-means centers and PC scores, respectively.

C. ANN Architecture

A multilayer perceptron with back-propagation learning was employed as the ANN (Figure 4). When ANN training

TABLE I

DATA USED IN ANN TRAINING, RECURSIVE FEATURE ELIMINATION (RFE), AND CLAMPING CRITERIA.

Data	ANN usage	Reduction method K-means	Reduction method PCA
Matrices	Train[a]	$\underset{A\times K}{\mathbf{M}}$	$\underset{A\times P}{\mathbf{F}}$
	RFE[b]	$\underset{G\times K}{\mathbf{Z}}$	$\underset{G\times P}{\mathbf{W}}$
	Prediction[c]	$\underset{A\times G}{\mathbf{E}}$	$\underset{A\times G}{\mathbf{E}}$
ANN clamping vectors	Train	$\mathbf{x}_a = \mathbf{m}_a$	$\mathbf{x}_a = \mathbf{f}_a$
	RFE	$\mathbf{x}_g = \mathbf{z}_g$	$\mathbf{x}_g = \mathbf{w}_g$
	Prediction	$\mathbf{x}_a = \mathbf{e}_a$	$\mathbf{x}_a = \mathbf{e}_a$

[a] Matrices **M** and **F** used for training ANN.
[b] Applied weights after last training epoch against matrices **Z** and **W** looping over all genes once to determine E_c^g and S_c^g.
[c] Class membership was predicted after retraining ANN with standardized expression matrix **E** for genes identified during RFE.
M is the matrix of k-means centers.
F is the matrix of PC scores.
Z is the matrix of k-means scores.
W is the matrix of PC score score coefficients.
E is the matrix of standardized expression.
A is the number of arrays, ($a = 1, 2, \ldots, A$) .
K is the number of k-means cluster centers, ($k = 1, 2, \ldots, K$) .
P is the number of PCs, ($p = 1, 2, \ldots, P$) .
G is the number of genes, ($g = 1, 2, \ldots, G$) .
During RFE, $G = 725$, whereas during prediction $G = 2, 4, 8, 16, 32, 64$.

was based on k-means centers, we clamped rows of the $A \times K$ **M** matrix of cluster centers to the input nodes (i.e., $\mathbf{x}_a = \mathbf{m}_a$, $a = 1, 2, \ldots, 32$). Whereas when ANN training was based on PCA, we clamped rows of the $A \times P$ **F** matrix of PC scores to the input nodes (i.e., $\mathbf{x}_a = \mathbf{f}_a$, $a = 1, 2, \ldots, 32$). The number of hidden nodes was equal to 40% of the number of input nodes. Connection weights w_{ij}^{ih} between the input and hidden layer and w_{jc}^{ho} between hidden and output layers were initialized with random uniform variates in the range [-0.5,0.5]. Node output v_j at the hidden layer were based on the logistic activation function $1/(1 + \exp(-u_j))$, where $u_j = \sum x_i w_{ij}^{ih}$, while node outputs $\hat{t}_c$ at the output layer were based on the softmax function $\exp(y_c)/\sum \exp(y_l)$, where $y_c = \sum v_j w_{jc}^{ho}$. Mean square error (MSE) for each sample was determined as $0.5 \sum_c (\hat{t}_c - t_c)^2$, where $\hat{t}_c$ is the predicted target probability of each sample being in class c and t_c is the true class probability of each sample, set to zero or unity. After each sample was transmitted through the ANN, back-propagation learning was performed sequentially updating weights one sample at a time. A total of 25 epochs was performed for each model, with 32 cycles (samples) per epoch. (Models using 5000 epochs were evaluated and neither increased overtraining or diversion of the gradient descent toward local maxima were observed). The ANN model fitting approach combined leave-one-out testing with 8-cross-validation. Arrays in the validation group were not used for training and only for prediction of class outcome, while the remaining samples were used for training the ANN. Last, we used a learning rate of $\epsilon = 0.5$, a steadily decreasing weight decay γ, and momentum of $\alpha = 0.5$.

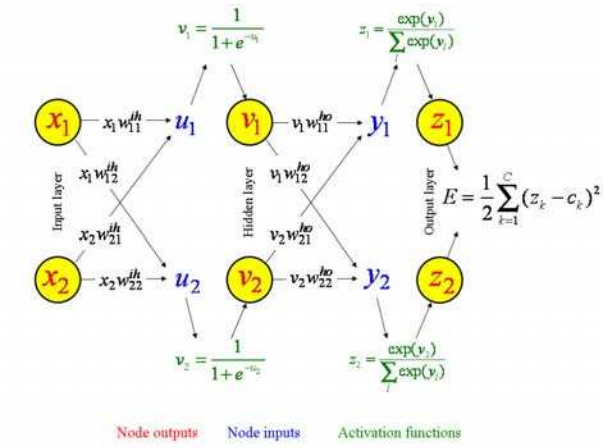


Fig. 4. Schematic of the artificial neural network (ANN) used.

D. ANN Recursive Feature Elimination

In order to gauge the influence of each gene on the classification, target outputs $\hat{t}_c^g$ for each gene were calculated during the last sweep of every model using the last known weights and setting the input nodes x_i equal to either the $1 \times K$ row vector of k-means scores $\mathbf{z}_g$ for each gene or the $1 \times P$ row vector $\mathbf{w}_g$ of PC score coefficients for each gene. It warrants noting that the ANN was not retrained here, but rather gene-specific values of $\hat{t}_c^g$ were determined by applying the last known weights to gene-specific row vectors of **Z** or **W**, which map the genes back to the original **M** and **F** matrices used for training.

a) Gene Selection with Maximum Sensitivity: The average gene-class-specific sensitivity [24] of each gene was determined as

$$S_c^g = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{t}_c^g}{\partial x_i} \tag{2}$$

where g is the gene, c is the class, n is the number of input nodes based on $n = K$ and $\mathbf{x}_g = \mathbf{z}_g$ if the ANN was trained with k-means centers based on **M**, or $n = P$ and $\mathbf{x}_g = \mathbf{w}_g$ if the ANN was trained with PC scores based on **F** (Table I). Class-specific sensitivities for each gene were summed over all 256 models and then sorted in descending order.

b) Gene Selection with Minimum Error: In addition to RFE based on sensitivity, we also calculated the gene-class-specific mean square error during the last sweep, using the recomputed values of $\hat{t}_c^g$ described above. Analogously, we derived lists of genes for which each class was represented equally by genes having the lowest gene-class-specific MSE. Recall that the predicted class target $\hat{t}_c^g$ for each gene is class specific; however, genes do not have a true class membership, i.e., t_c^g, so we calculated error as $E_c^g = \sum_j^C 0.5(\hat{t}_c^g - I(j)^2$, where $I(j)$ is one if $j = c$ and zero if $j \neq c$.

E. Generating Lists of Selected Genes

A modular approach was employed for generating the list of genes identified during RFE. Lists were divided uniformly into genes that best discriminated each outcome class, depending on whether the selection criterion was minimum gene-class-specific MSE or maximum gene-class-specific sensitivity. The total number of genes in a list was based on powers of 2 multiplied by the number of classes, such that the list was uniformly loaded with genes that best discriminated each class. As an example, in this two class study, gene selection lists contained a total of 2, 4, 8, 16, 32, or 64 genes with half of the genes having the total maximum sensitivity (or least error) for the poor outcome and half having the maximum sensitivity (or least error) for the good outcome class.

F. ANN Training with Selected Gene Expression Profiles

After recursive feature identification, we trained the ANN models with the actual standardized values of expression for each list of 2, 4, 8, 16, 32, or 64 genes identified. Expression for each gene was standardized over the arrays (using array-specific mean and s.d.) ANN clamping involved use of the rows of the standardized expression matrix $\mathbf{E}$ such that $\mathbf{x}_a = \mathbf{e}_a$ (Table I). The number of hidden nodes was equal to 40% of the number of input nodes. For example, for 64 genes (features) and 2 outcome classes a 64-26-2 network was employed. During runs with actual gene expression profiles, we assessed accuracy and, the proportion of between-gene correlation coefficients that were significant ($p \leq 0.01$).

G. ANN Pseudocode

The following provides the ANN algorithm used:

1) **for** *leaveout* $\leftarrow$ 1 to #Arrays
2) Randomly assign each array into O equal partitions $(o = 1, 2, ..., O)$
3) **for** $o \leftarrow 1$ to O
4) Initialize $\{w_{ij}^{ih}\}$ and $\{w_{jc}^{ho}\}$ in the range [-0.5,0.5].
5) $\Delta w_{ij}^{ih} \leftarrow 0, \forall i, j$, $\Delta w_{jc}^{ho} \leftarrow 0, \forall j, c$.
6) **for** *sweep* $\leftarrow$ 1 to #Epochs
7) $t \leftarrow 0$
8) **for** *array* $\leftarrow$ 1 to #Arrays
   - $u_j = \sum_{i=1}^{I} x_i w_{ij}^{ih}, \forall j$.
   - $v_j = 1/(1 + e^{-u_j}), \forall j$.
   - $y_c = \sum_{j=1}^{H} v_j w_{jc}^{ho}, \forall c$
   - $\hat{t}_c = \exp(y_c)/\{\sum_{l=1} \exp(y_l)\}, \forall c$
   - $E = \frac{1}{2}\sum_{c=1}^{C}(\hat{t}_c - t_c)^2$
   - $E_c = \frac{1}{2}(\hat{t}_c - t_c)^2$
   - **if** *array* $\notin o$ **then** 'training array
     $t = t + 1$ 'increment cycle
     $\Delta w_{jc}^{ho}(t) = -\varepsilon\gamma(\partial E/\partial w_{jc}^{ho}) + \alpha\Delta w_{jc}^{ho}(t-1), \forall j, c$.
     $w_{jc}^{ho}(t) = w_{jc}^{ho}(t-1) + \Delta w_{jc}^{ho}(t), \forall j, c$.
     $\Delta w_{ij}^{ih}(t) = -\varepsilon\gamma(\partial E/\partial w_{ij}^{ih}) + \alpha\Delta w_{ij}^{ih}(t-1), \forall i, j$.
     $w_{ij}^{ih}(t) = w_{ij}^{ih}(t-1) + \Delta w_{ij}^{ih}(t), \forall i, j$.
   - **endif**
   - **if** *sweep* = #Epochs **then** 'last sweep
     if *array* $\in o$ **then** 'validation array
     $pdf(c(array))_{diagnosis} = E_{c(array)}$
     endif
     if *array* = *leaveout* **then** 'left out array
     $pdf(array)_{prediction} = E_{c(array)}$
     endif
   - **endif**
9) **Next** *array*
10) **Next** *sweep*
11) Call geneidentification(S_c^g).
12) **Next** o
13) **Next** *leaveout*

III. Results

Figures 5 and 6 show the MSE for ANN training based on standardized expression profiles for the top 4 genes and 16 genes, respectively, identified after k-means dimensional reduction and maximum sensitivity used as a criterion for RFE. In Figures 5 and 6, red denotes training MSE while blue denotes validation MSE. In Figure 5, one can notice the overtraining that occurred as a result of ANN training with expression profiles for only 4 genes. When the ANN was trained with expression profiles of the top 16 genes (Figure 6), however, there was no apparent overtraining for the 25 epochs used.

Tables II and III list diagnostic sensitivity and specificity results of ANN modeling for the poor and good outcome class predictions, respectively. Recall that *sensitivity* is defined as the probability of correctly predicting that an array is in class A given that the array is truly from diagnostic class A. *Specificity* is defined as the probability of correctly assigning an array to a class other than A given that the array is truly not in class A. For the calibration run using the k-means centers and PC scores from all 725 probes, sensitivity and specificity ranged from 90-100% for both outcome groups. For runs using standardized expression for the top genes identified, at 2 genes and more, sensitivity and specificity steadily ramped up from 70-80% for 2 genes to the greatest values in the 90% range for 64 genes. Interestingly, for k-means cluster analysis dimensional reduction, the maximum sensitivity and specificity for both outcomes was observed for 16-gene models. However, for PCA dimensional reduction, the greatest values of sensitivity and specificity were observed among 64-gene models. These results show that, for the training data considered, use of k-means cluster analysis for data pre-processing, followed by cross-validation and leave-one-out testing can provide remarkably high screening diagnostics exceeding 95%. The 16 genes identified after k-means dimensional reduction and maximum sensitivity for RFE that resulted in a sensitivity of 1.0 (in Table II) were *DGCR5, FLJ10618, RIS1, PRO1855, ABCB9, AK057203, GOLGA5, HARS, AK024152, HEP27, PPIA, SNRPF, SULT1A3, SECTM1, EIF4EBP1*, and *S71435*. These are also highlighted in bold in Table IV in the 64-gene model.

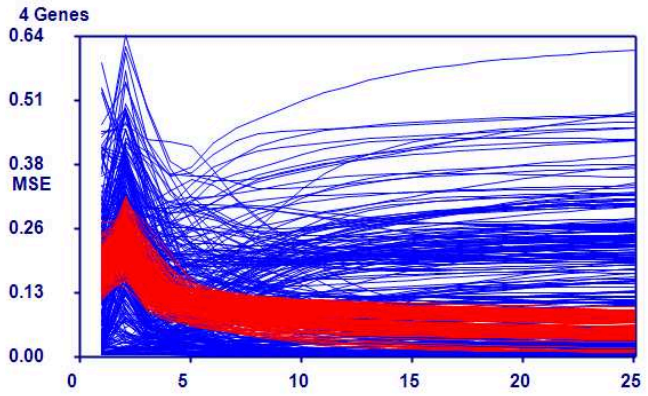


Fig. 5. MSE for 256 ANN models as a function of training epoch for a 4-gene model based on k-means dimensional reduction and maximum sensitivity for RFE. Red denotes training MSE while blue denotes validation MSE.

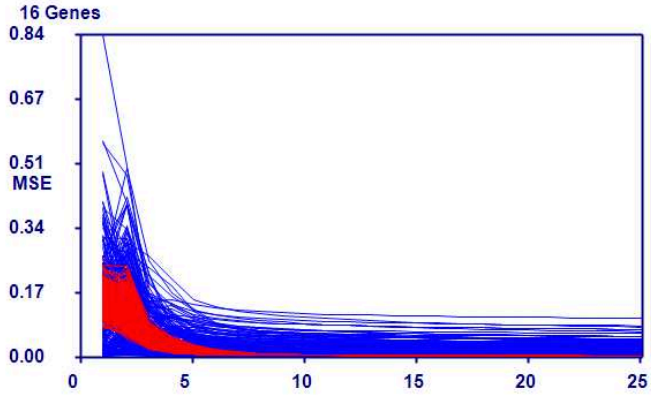


Fig. 6. MSE for 256 ANN models as a function of training epoch for a 16-gene model based on k-means dimensional reduction and maximum sensitivity for RFE. Red denotes training MSE while blue denotes validation MSE.

Tables IV and V list the top 64 genes selected by RFE for the 64-gene models for dimensional reduction with k-means cluster analysis and PCA, respectively. It was interesting to learn that the 64 genes listed were not always the most significant genes based on the Mann-Whitney U test used for the gene ranking. Although not shown, the majority of the genes identified were among the least significant (rank >300). This reflects the generalizability of ANN models for which there is a preference to select not the most significant genes, but rather those which simultaneously maximize the screening diagnostics (sensitivity and specificity). The genes listed in Tables IV and V became available at the time of writing, so they are now being mined for their roles in cancer (i.e., prostate), GO groups, PFAM familes, secretion signals, transcription factor binding sites in promoters, and protein interactions,

TABLE II

Diagnostic sensitivity of class prediction based on ANN training with standardized expression for selected genes.

k-means				
	$\max\{S_c^g\}$		$\min\{E_c^g\}$	
#Genes	Poor	Good	Poor	Good
725[a]	1.000	0.985	1.000	0.985
2	0.671	0.786	0.852	0.869
4	0.860	0.833	0.908	0.879
8	0.961	0.877	0.903	0.886
16[c]	**1.000**	**0.991**	0.959	0.884
32	0.985	0.985	0.959	0.898
64	0.961	0.960	0.991	0.924
PCA				
	$\max\{S_c^g\}$		$\min\{E_c^g\}$	
#Genes	Poor	Good	Poor	Good
725[b]	0.940	0.867	0.940	0.867
2	0.708	0.816	0.746	0.687
4	0.837	0.787	0.826	0.786
8	0.774	0.770	0.768	0.763
16	0.843	0.825	0.768	0.782
32	0.869	0.882	0.869	0.863
64	0.914	0.903	0.899	0.888

[a] ANN trained with **M** matrix of k-means centers.
[b] ANN trained with **F** matrix of PC scores.
[c] 16 genes are listed in Table IV.

etc.

Figure 7 illustrates the significant positive and negative between-gene correlation of expression for the 64-gene models. Red represents significant ($p < 0.01$) positive correlation coefficients, while blue denotes significant ($p < 0.01$) negative correlation coefficients. Panel 4A for k-means dimensional reduction and $\min\{E_c^g\}$ shows that 0.444 of the correlation coefficients were significantly positive, panel 4B for k-means and $\max\{S_c^g\}$ shows that only 0.185 of the correlation coefficients were significantly positive, panel 4C for PCA and $\min\{E_c^g\}$ reflects that 0.295 of the correlation coefficients were significantly positive, and panel 4D for PCA and $\max\{S_c^g\}$ illustrates that 0.287 of the coefficients were significantly positive. Overall, one can note the reduced significant positive correlation for RFE with maximum sensitivity (panels 4B and 4D), which is accentuated for k-means dimensional reduction (panel 4B). This is in good agreement with results obtained in our previous work using simulated and empirical expression data from the public domain [24].

IV. Discussion and Conclusion

The purpose of this study was to compare diagnostic screening results (sensitivity and specificity) and betweengene correlation across dimensional reduction and RFE methods used. The ANN analysis results obtained demonstrate that diagnostic sensitivity and specificity in the range 99-100% was obtained when using k-means dimensional reduction and maximum sensitivity for RFE for a 16-gene model. On the other hand, for PCA dimensional reduction, diagnostic screening results were in the 70-80%

TABLE III

Diagnostic specificity of class prediction based on ANN training with standardized expression for selected genes.

k-means				
	max$\{S_c^g\}$		min$\{E_c^g\}$	
#Genes	Poor	Good	Poor	Good
725[a]	0.985	1.000	0.985	1.000
2	0.786	0.671	0.869	0.852
4	0.833	0.860	0.879	0.908
8	0.877	0.961	0.886	0.903
16	**0.991**	**1.000**	0.884	0.959
32	0.985	0.985	0.898	0.959
64	0.960	0.961	0.924	0.991
PCA				
	max$\{S_c^g\}$		min$\{E_c^g\}$	
#Genes	Poor	Good	Poor	Good
725[b]	0.867	0.940	0.867	0.940
2	0.816	0.708	0.687	0.746
4	0.787	0.837	0.786	0.826
8	0.770	0.774	0.763	0.768
16	0.825	0.843	0.782	0.768
32	0.882	0.869	0.863	0.869
64	0.903	0.914	0.888	0.899

[a] ANN trained with **M** matrix of k-means centers.
[b] ANN trained with **F** matrix of PC scores.

range for 16 genes. In addition, dimensional reduction with k-means cluster analysis and RFE with maximum sensitivity resulted in the least amount of significant positive between-gene correlation.

We are currently trying to understand the biological relevance of the 64 genes identified with the various methods. The majority of this effort involves GO mining and development of transcription factor networks and review by biologists to determine any differences in the computational results. We will eventually compare results with support vector machines, and soft computing methods, which provide alternative approaches for optimizing decision boundaries between classes.

Acknowledgments

Research was supported by NCI grant K22-CA10028-02, a Baylor Prostate Cancer SPORE (P50CA058204) pilot project, and by use of the facilities of the Michael E. DeBakey Dept. of Veterans Affairs Medical Center. The technical assistance of Rebecca Penland and Shantu Dixit is gratefully acknowledged.

References

[1] Jemal A, Tiwari RC, Murray T, et al. Cancer statistics, 2004. CA Cancer J Clin 2004;54:8-29.
[2] Aihara M, Wheeler TM, Ohori M, Scardino PT. Heterogeneity of prostate cancer in radical prostatectomy specimens. Urology 1994;43:60-6.
[3] Gerber GS, Thisted RA, Scardino PT, et al. Results of radical prostatectomy in men with clinically localized prostate cancer. JAMA 1996;276:615-9.

TABLE IV

Genes identified in 64-gene models for k-means cluster analysis dimensional reduction. Genes in bold resulted in greatest diagnostic sensitivity and specificity among all other gene models, see Table II.

max$\{S_c^g\}$	min$\{E_c^g\}$
DGCR5	*BYSL*
FLJ10618	*ADM*
RIS1	*EIF4EL3*
PRO1855	*PPT2*
ABCB9	*BC012900*
AK057203	*PIP5K1C*
GOLGA5	*AK057233*
HARS	*PFKFB1*
L33988	*HTR1A*
AK022045	*KIAA0575*
AJ420571	*FLJ10900*
GLRX2	*AF070574*
CYC1	*ALPP*
DKFZp434N0650	*KIAA0878*
IRF2	*PRKCBP1*
TMEPAI	*LOC56899*
FLJ11585	*RAD51C*
GNG8	*C20orf67*
ZNF302	*AK055723*
SCRIB	*SLC11A1*
NPHP1	*FLJ20040*
SCAD-SRL	*AK022635*
KIAA1522	*IL3*
FLJ10986	*ASK*
MGC14833	*AK023312*
PIG6	*MPDU1*
NM_053284	*AJ420416*
CABP2	*CCNB2*
AK056890	*PRSC*
DUSP10	*MYBL2*
FLJ10521	*BICD1*
FLJ20778	*AK002000*
AK024152	*DKFZp762P2111*
HEP27	*CAB56184*
PPIA	*PIGR*
SNRPF	*RASA2*
SULT1A3	*FLJ20003*
SECTM1	*FABP7*
EIF4EBP1	*AK025902*
S71435	*MTCP1*
APOB	*DMWD*
DHPS	*AK025806*
AK026657	*BRF2*
SNRPG	*SYN2*
TNFAIP3	*NR6A1*
AK027185	*AK026984*
MGC3032	*SNX4*
COX7A2	*FLJ20081*
K-ALPHA-1	*LOC57862*
MPDU1	*CACNG5*
EIF4EL3	*FLJ00052*
BMPR1A	*AB051444*
AK055589	*OT7T022*
AK057233	*PVALB*
KIAA1093	*AC008050*
PRO1073	*CNNM3*
AF052181	*DDAH2*
AK026439	*FLJ10043*
U79289	*AL136753*
AJ420528	*AK056119*
FLJ10900	*PP15*
BYSL	*X55777*
PIP5K1C	*PIPPIN*
ADM	*FLJ14442*

TABLE V

GENES IDENTIFIED IN 64-GENE MODELS FOR PCA DIMENSIONAL REDUCTION.

$\max\{S_c^g\}$	$\min\{E_c^g\}$
PRO1073	*IRF2*
POLR2J	*DKFZP564P1916*
AK026657	*SRP54*
SULT1A3	*BC007749*
MGC5356	*TDPGD*
SAFB	*GLRX2*
AK024289	*FLJ10675*
SNRPF	*MMP7*
DKFZp762O076	*AK027155*
SNRPG	*AK023744*
PGLYRP	*PRO0297*
AK023312	*COVA1*
BMPR1A	*AL133086*
AK026704	*HOXA7*
FLJ13942	*ZNF197*
DRIL1	*EEA1*
MRPL17	*PHYHIP*
BAZ1A	*RAB4B*
AK055603	*CGI-57*
MGC15730	*NRF1*
DGCR5	*GRP58*
FLJ20313	*ZNF275*
PSMD3	*RNF28*
DHPS	*FLJ23306*
AL133050	*PP1044*
AK022294	*BC002782*
SAP18	*AK022134*
AK055203	*AK000435*
BC016154	*AK023945*
WNT7A	*KIFAP3*
FLJ20442	*GGT1*
KRTHA4	*FLJ14855*
IRF2	*POLR2J*
DKFZP564P1916	*PRO1073*
SRP54	*SULT1A3*
BC007749	*AK026657*
TDPGD	*SAFB*
GLRX2	*AK024289*
MMP7	*SNRPF*
FLJ10675	*SNRPG*
AK027155	*MGC5356*
AK023744	*DKFZp762O076*
HOXA7	*AK023312*
EEA1	*PGLYRP*
PRO0297	*BMPR1A*
AL133086	*AK055603*
COVA1	*AK026704*
ZNF197	*FLJ20313*
NRF1	*MRPL17*
GRP58	*FLJ13942*
RAB4B	*DHPS*
PHYHIP	*DRIL1*
FLJ23306	*BAZ1A*
KIFAP3	*AK055203*
CGI-57	*PGCP*
ZNF275	*KPTN*
BC002782	*MGC15730*
AK000435	*BC016154*
AK022134	*AK022294*
GGT1	*FLJ20442*
RNF28	*WNT7A*
FLJ14855	*PSMD3*
PP1044	*AL133050*
AK023945	*DGCR5*

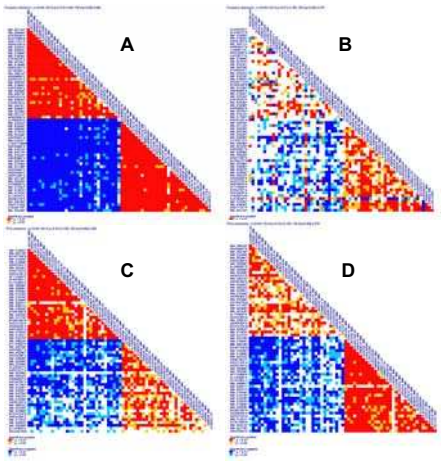


Fig. 7. Between-gene correlation for 64-gene models. Red represents significant ($p < 0$ .01) positive correlation coefficients, while blue denotes significant ($p < 0$ .01) negative correlation coefficients. (A) k-means dimensional reduction and RFE with $\min\{E_c^g\}$, (B) K-means and RFE with $\max\{S_c^g\}$, (C) PCA and RFE with $\min\{E_c^g\}$, (D) PCA and RFE with $\max\{S_c^g\}$.

[4] Lukka H, Warde P, Pickles T, et al. Controversies in prostate cancer radiotherapy: consensus development. Can J Urol 2001;8:1314-22.

[5] Porter AT, Blasko JC, Grimm PD, et al. Brachytherapy for prostate cancer. CA Cancer J Clin 1995;45:165-78.

[6] Bogers JA, van der Maazen RW, Visser AG. Conformal photon-beam radiotherapy of prostate carcinoma. Eur Urol 2002;41:515-22.

[7] Dipetrillo TA. Three-dimensional conformal radiotherapy for early stage prostatic cancer: techniques, outcomes, and possible pitfalls. Surg Oncol Clin N Am 2000;9:435-53, viii.

[8] Augustin H, Hammerer PG. Disease recurrence after radical prostatectomy. Contemporary diagnostic and therapeutical strategies. Minerva Urol Nefrol 2003;55:251-61.

[9] Scheepe JR, Vis AN, Mickisch GH. Recurrence after radical prostatectomy for organ-confined prostate cancer. Onkologie 2003;26:330-3.

[10] Ward JF, Blute ML, Slezak J, et al. The long-term clinical impact of biochemical recurrence of prostate cancer 5 or more years after radical prostatectomy. J Urol 2003;170:1872-6.

[11] Hull GW, Rabbani F, Abbas F, et al. Cancer control with radical prostatectomy alone in 1,000 consecutive patients. J Urol 2002;167:528-34.

[12] Pound CR, Partin AW, Eisenberger MA, et al. Natural history of progression after PSA elevation following radical prostatectomy. JAMA 1999;281:1591-7.

[13] Lapointe J, Li C, Higgins JP, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci U S A 2004;101:811-6.

[14] Singh D, Febbo PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002;1:203-9.

[15] Glinsky GV, Glinskii AB, Stephenson AJ, et al. Gene expression profiling predicts clinical outcome of prostate cancer. J Clin Invest 2004;113:913-23.

[16] Yu YP, Landsittel D, Jing L, et al. Gene expression alterations in prostate cancer predicting tumor aggression

and preceding development of malignancy. J Clin Oncol 2004;22:2790-9.

[17] Khan J, Wei JS, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7:673-9.

[18] Wei JS, Greer BT, Westermann F, et al. Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma. Cancer Res 2004;64:6883-91.

[19] Ando T, Suguro M, Hanai T, et al. Fuzzy neural network applied to gene expression profiling for predicting the prognosis of diffuse large B-cell lymphoma. Jpn J Cancer Res 2002;93:1207-12.

[20] Bicciato S, Pandin M, Didone G, Di Bello C. Pattern identification and classification in gene expression data using an autoassociative neural network model. Biotechnol Bioeng 2003;81:594-606.

[21] O'Neill MC, Song L. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. BMC Bioinformatics 2003;4:13.

[22] Liu B, Cui Q, Jiang T, Ma S. A combinational feature selection and ensemble neural network method for classification of gene expression data. BMC Bioinformatics 2004;5:136.

[23] Hand DJ, Mannila H, Smyth P. Principles of Data Mining (Adaptive Computation and Machine Learning). Cambridge(MA): MIT Press, 2001.

[24] Peterson LE, Coleman MA. Comparison of gene identification based on artificial neural network pre-processing with k-means cluster and principal component analysis. Proceedings of the WILF 2005 Workshop, Milan(Crema), Italy. Lecture Notes in Computer Science, Springer, Berlin, 2005.