Comparison of Gene Identification Based on Artificial Neural Network Pre-processing with k-Means Cluster and Principal Component Analysis

Leif E. Peterson[1] and Matthew A. Coleman[2]

[1] Baylor College of Medicine, Houston, TX 77030, USA
peterson@bcm.tmc.edu
http://www.chipst2c.org
[2] Lawrence Livermore National Laboratory, Livermore, CA 94550, USA

Abstract. A combination of gene ranking, dimensional reduction, and recursive feature elimination (RFE) using a BP-MLP artificial neural network (ANN) was used to select genes for DNA microarray classification. Use of k-means cluster analysis for dimensional reduction and maximum sensitivity for RFE resulted in 64-gene models with fewer invariant and correlated features when compared with PCA and mimimum error. In conclusion, k-means cluster analysis and sensitivity may be better suited for classifying diseases for which gene expression is more strongly influenced by pathway heterogeneity.

1 Introduction

Artificial neural networks (ANNs) have been applied to DNA microarray data through several approaches. Tarca et al used ANNs to normalize cDNA microarray data and demonstrated a reduction in both intensity-dependent bias and spatial-dependent bias[1]. The agreement between regulatory motifs and functional classes of *Saccharomyces cerevisiae* genes in clusters based on Euclidean distance, correlation, and mutual information was found to be lower than ANN-derived clusters[2]. Using expression data for cardiovascular disease, Tham et al reported that an ANN approach provided promising prediction results[3]. The remaining clinical papers on ANNs focused on diagnostic classification of several types of cancer such as leukemia, lymphoma, lung cancer, prostate cancer, and various neurological malignancies [4-9].

The goal of this paper was to assess correlation and differential expression among features identified through a combination of methods involving gene ranking, dimensional reduction, and recursive feature elimination (RFE). Comparisons are provided describing the amount of between-gene correlation in 64-gene models as a function of dimensional reduction and RFE methods. Also provided is the proportion of genes among the 64-gene models with significant between-class differential expression.

I. Bloch, A. Petrosino, and A.G.B. Tettamanzi (Eds.): WILF 2005, LNAI 3849, pp. 267–276, 2006.

2 Methods

2.1 Simulated Data Set

Let the matrix **E** of gene expression profiles have dimension $G \times A$, where G is the number of genes ($g = 1, 2, \ldots, G$) and A is the number of microarrays ($a = 1, 2, \ldots, A$). A simulated data set with 400 genes and 20 arrays was generated with expression values distributed normally as N(μ, σ^2) as shown in Table 1. Symmetry in the simulated expression values was preserved among the 2 classes in order to prevent bias in the sensitivity for a particular class.

Table 1. Description of 400 simulated genes for 20 arrays

Simulated genes				
	Class A(10 arrays)		*Class B(10 arrays)*	
# Genes	*Odd arrays*	*Even arrays*	*Odd arrays*	*Even arrays*
40	N(0,1)	N(0,1)	N(0,1)	N(0,1)
20	N(5,1)	N(5,1)	N(0,1)	N(0,1)
20	N(0,1)	N(0,1)	N(5,1)	N(5,1)
20	N(5,1)	N(5,1)	N(-5,1)	N(-5,1)
20	N(-5,1)	N(-5,1)	N(5,1)	N(5,1)
20	N(5,1)	N(0,1)	N(0,1)	N(0,1)
20	N(0,1)	N(0,1)	N(5,1)	N(0,1)
20	N(5,1)	N(-5,1)	N(0,1)	N(0,1)
20	N(0,1)	N(0,1)	N(5,1)	N(-5,1)
20	N(5,1)	N(-5,1)	N(5,1)	N(0,1)
20	N(5,1)	N(0,1)	N(5,1)	N(-5,1)
20	N(5,1)	N(-5,1)	N(5,1)	N(-5,1)
20	N(2.5,1)	N(0,1)	N(0,1)	N(0,1)
20	N(0,1)	N(0,1)	N(2.5,1)	N(0,1)
20	N(2.5,1)	N(-2.5,1)	N(0,1)	N(0,1)
20	N(0,1)	N(0,1)	N(2.5,1)	N(-2.5,1)
20	N(2.5,1)	N(-2.5,1)	N(2.5,1)	N(0,1)
20	N(2.5,1)	N(0,1)	N(2.5,1)	N(-2.5,1)
20	N(2.5,1)	N(-2.5,1)	N(2.5,1)	N(-2.5,1)

2.2 Empirical Data Sets

We used two empirical data sets available in the public domain. The first was published by Hedenfalk et al [10] on *BRCA1* and *BRCA2* mutations with 3170 genes and 15 arrays comprising 2 classes (7 arrays for *BRCA1* and 8 arrays for *BRCA2*). The second was published by Khan et al [9] on childhood small round blue-cell tumors (SRBCT) with 2308 genes and 63 arrays comprising 4 classes (23 arrays for EWS-Ewing Sarcoma, 8 arrays for BL-Burkitt lymphoma, 12 arrays for NB-neuroblastoma, and 20 arrays for RMS-rhabdomyosarcoma).

2.3 Gene Ranking

We applied non-parametric independent k-sample statistical tests and ranked genes based on their significance level. For the 2-class simulated and Hedenfalk et al data sets, we applied the Mann-Whitney test to rank genes based on significance. The Mann-Whitney test approximates the Gini diversity index commonly

used for feature selection[11]. While all of the 400 simulated genes were used, we applied a cutoff criterion of $p \leq 0.2$ for the 3170 original genes in the Hedenfalk et al data set and identified 967 genes. For the 4-class Khan et al data set with 2308 original genes, we used the independent k-sample Kruskal-Wallis ANOVA test to rank genes and applied a cutoff $p \leq 0.01$. This led to 898 gene expression profiles. Tail probabilities for parametric t-tests applied to the Hedenfalk et al data resulted in 1191 genes with $p \leq 0.20$ and 920 genes for F-tests applied to the Khan et al data for which $p \leq 0.01$. Parametric test results were not used for gene ranking, but stored for bookkeeping.

2.4 Dimension Reduction with K-Means Cluster Analysis

In addition to gene ranking and use of p-value cutoffs, we applied k-means clustering and principal components analysis (PCA) for dimension reduction in order to minimize effects from the curse of dimensionality[12]. For k-means clustering, let k (k=1,2,...,K) be the the kth cluster of a clustering, and K the total number of clusters. The optimal value of K is determined by cycling through values of $K = 2, 3, ..., \sqrt{G}$. This is performed as follows. For K clusters, the total *within-cluster sum-of-squares* is

$$SSW(K) = \sum_{k=1}^{K} \sum_{g=1}^{G_k} \|\mathbf{x}_{gk} - \mathbf{m}_k\|, \tag{1}$$

where $\mathbf{x}_{gk}$ is the row vector containing expression values for gene g in cluster k over the A arrays and $\mathbf{m}_k$ is the mean vector for G_k genes in cluster k, and $\|.\|$ is the Euclidean distance. For the same K clusters, the smallest *between-cluster distance* is

$$d(K) = \min_{1 \leq k < l \leq K} \|\mathbf{m}_k - \mathbf{m}_l\|, \tag{2}$$

and the score function for a set of K clusters is

$$S_K = \frac{d(K)}{SSW(K)}. \tag{3}$$

After evaluating the score function S_K for values of K ranging from 2 to $\sqrt{G}$, the optimal value of K is

$$K_{opt} = \max_{2 \leq K \leq \sqrt{G}} \{S_K\}. \tag{4}$$

Once K_{opt} is determined, the k-means algorithm is rerun using K_{opt} clusters. K-means clustering results in a $A \times K$ $\mathbf{M}$ matrix of *k-means centers*. For each gene, determine the k-means score which maps the gene back to the center k as

$$z_{gk} = \frac{\|\mathbf{x}_g - \mathbf{m}_k\| - \mu_k}{\sigma_k} \qquad k = 1, 2, \ldots, K, \tag{5}$$

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 $G \times K$ $\mathbf{Z}$ matrix of *k-means scores*. Since the scores are standard normal distributed, the bulk of scores will be centered around zero and genes with the smallest or greatest distance from the cluster center will yield greater scores. For the simulated data set, we identified 12 centers (i.e., K_{opt}=12). For the Hedenfalk et al breast cancer data, we identified 30 centers, whereas for the Khan et al SRBCT data, we identified 28 centers.

2.5 Dimension Reduction with PCA

During PCA, the top 10 eigenvalues were extracted from the $G \times G$ correlation matrix $\mathbf{R}$. The array by component $(A \times P)$ $\mathbf{F}$ matrix of *PC scores* is determined with the matrix of standardized expression values (standardized with mean and s.d. over the genes) as follows

$$\underset{A\times P}{\mathbf{F}} = \underset{A\times G}{\mathbf{Z}}\,\underset{G\times P}{\mathbf{W}}. \tag{6}$$

In order to map genes back to the arrays via the principal components, the matrix of *PC score coefficients* was obtained using the matrix operation

$$\underset{G\times P}{\mathbf{W}} = \underset{G\times P}{\mathbf{L}}\,\underset{P\times P}{(\mathbf{L}'\mathbf{L})^{-1}}, \tag{7}$$

where $\mathbf{L}$ is the loading matrix reflecting the correlation between each gene expression profile and the extracted PC scores. The top 10 PC's were always extracted from the gene by gene correlation matrix and used for training. Orthogonal rotations were not performed.

2.6 ANN Training During Recursive Feature Elimination

Recursive feature elimination (RFE) was based on a BP-MLP ANN with one hidden layer. The ANNChip computer program (http://www.chipst2c.org) was used for RFE and included 8-fold cross-validation and leave-one-out testing where arrays were randomly assigned to 8 validation groups. Each validation group was selected singly resulting in a single ANN model in which the remaining 7/8 of arrays were used for training. During leave-one-out testing, array 1 was left out of models 1-8, array 2 left out of models 9-16, etc., so that each array was left out during 8 models. Table 2 summarizes the input data sets with the number of samples and genes, the reduction methods and derived matrices used to feed the ANN during RFE, and the total number of models used based on 8-fold cross-validation with leave-one-out testing.

Selection of Genes Based on Maximum Sensitivity. 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

Table 2. ANN training during recursive feature elimination (RFE) and ANN model summary for 8-fold cross-validation with leave-one-out testing

Data	*Reduction method*	*Samples*	*Genes*	*Training matrix*[a]	*Network size*	*ANN models*[b]
Simulated	k-means	20	400	$\mathbf{M}_{20\times12}$	12-5-2	160
	PCA	20	400	$\mathbf{F}_{20\times10}$	10-4-2	160
Hedenfalk et al	k-means	15	967	$\mathbf{M}_{15\times30}$	30-12-2	120
	PCA	15	967	$\mathbf{F}_{15\times10}$	10-4-2	120
Khan et al	k-means	63	898	$\mathbf{M}_{63\times28}$	28-11-4	504
	PCA	63	898	$\mathbf{F}_{63\times10}$	10-4-2	504

[a] Matrices from reduction results were used for training the ANN during RFE.
[b] Number of models equal to 8 validation groups times number of samples.

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 $\mathbf{Z}$ or $\mathbf{W}$, which map the genes back to the original $\mathbf{M}$ and $\mathbf{F}$ matrices used for training. The average gene-class-specific sensitivity of each gene was then determined as

$$S_c^g = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \hat{t}_c^g}{\partial x_i} \tag{8}$$

where g is the gene, c is the class, n is the number of input nodes based on $n = K$ and $\mathbf{x} = \mathbf{z}_g$ if the ANN was trained with k-means centers based on $\mathbf{M}$, or $n = P$ and $\mathbf{x} = \mathbf{w}_g$ if the ANN was trained with PC scores based on $\mathbf{F}$. The partial derivative $\partial \hat{t}_c^g / \partial x_i$ is determined via the chain rule, by first differentiating $\hat{t}_c^g$ w.r.t. hidden layer outputs, v_j, and then input row values, x_i, given by

$$\begin{aligned}
\frac{\partial \hat{t}_c^g}{\partial x_i} &= \sum_j \frac{\partial \hat{t}_c^g}{\partial v_j}\frac{\partial v_j}{\partial x_i} \\
&= \sum_j \frac{d\hat{t}_c^g}{dy_c}\frac{\partial y_c}{\partial v_j}\frac{dv_j}{du_j}\frac{\partial u_j}{\partial x_i} \\
&= \sum_j \left(\frac{\exp(y_c)\left(\sum_l \exp(y_l) - \exp(y_c)\right)}{\left(\sum_l \exp(y_l)\right)^2} w_{jc}^{ho} \frac{e^{-u_j}}{(1+e^{-u_j})^2} w_{ij}^{ih} \right)
\end{aligned} \tag{9}$$

Class-specific sensitivities for each gene were summed over all models and then sorted in descending order. Genes at the top of the sort were selected as the best predictors based on gene-class-specific sensitivity. The list of genes was divided equally into genes with the greatest sensitivity for discriminating each class.

For example, for a list of 8 genes and 2 classes, the 4 genes with the greatest sensitivity for discriminating class 1 were used along with the 4 genes with the greatest sensitivity for discriminating class 2.

Selection of Genes Based on Minimum Error. In addition to RFE based on sensitivity, we also calculated the gene-class-specific mean square error during the last sweep, $E_c^g = 0.5(\hat{t}_c^g - t_c)^2$, 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.

2.7 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.

2.8 ANN Training with Selected Gene Expression Profiles

After recursive feature identification, we trained the ANN models with the actual standardized values of expression for the identified genes. For example, for 64 genes (features) and 2 outcome classes a 64-26-2 network was employed, where the number of hidden nodes is equal to 40% of the number of input

Table 3. ANN training input using standardized expression profiles of genes selected during recursive feature elimination (RFE). 8-fold cross-validation with leave-one-out testing used.

Data	*Reduction method*	*RFE method*	*Samples*	*Genes*(n)	*Training matrix*[a]	*Network size*	*ANN models*[b]
Simulated	k-means	$\min(E_c^g)$	20	2,4,8,16,32,64	$\mathbf{E}_{20\times n}$	n-0.4n-2	160
		$\max(S_c^g)$	20	2,4,8,16,32,64	$\mathbf{E}_{20\times n}$	n-0.4n-2	160
	PCA	$\min(E_c^g)$	20	2,4,8,16,32,64	$\mathbf{E}_{20\times n}$	n-0.4n-2	160
		$\max(S_c^g)$	20	2,4,8,16,32,64	$\mathbf{E}_{20\times n}$	n-0.4n-2	160
Hedenfalk et al	k-means	$\min(E_c^g)$	15	2,4,8,16,32,64	$\mathbf{E}_{15\times n}$	n-0.4n-2	120
		$\max(S_c^g)$	15	2,4,8,16,32,64	$\mathbf{E}_{15\times n}$	n-0.4n-2	120
	PCA	$\min(E_c^g)$	15	2,4,8,16,32,64	$\mathbf{E}_{15\times n}$	n-0.4n-2	120
		$\max(S_c^g)$	15	2,4,8,16,32,64	$\mathbf{E}_{15\times n}$	n-0.4n-2	120
Khan et al	k-means	$\min(E_c^g)$	63	4,8,16,32,64	$\mathbf{E}_{63\times n}$	n-0.4n-4	504
		$\max(S_c^g)$	63	4,8,16,32,64	$\mathbf{E}_{63\times n}$	n-0.4n-4	504
	PCA	$\min(E_c^g)$	63	4,8,16,32,64	$\mathbf{E}_{63\times n}$	n-0.4n-4	504
		$\max(S_c^g)$	63	4,8,16,32,64	$\mathbf{E}_{63\times n}$	n-0.4n-4	504

[a] Training matrix of standardized gene expression $\mathbf{E}$ based on genes sorted by RFE method.
[b] Number of models equal to 8 validation groups times number of samples.

nodes. During runs with actual gene expression profiles, we assessed accuracy, the proportion of between-gene correlation coefficients that were significant ($p \leq 0.01$ and $p \leq 0.05$), and the proportion of genes in the list that had significant parametric test statistics ($p \leq 0.05$) during the original gene ranking calculations. Table 3 lists the gene expression data used for training the ANN.

3 Results and Discussion

The choice for using non-parametric Mann-Whitney U and Kruskal-Wallis tests for gene ranking should have a minimal effect on the observed results. Li et al assessed the effect of 8 different feature selection statistics on SVM outcome and determined that for more than 150 microarray-based genes the variation in performance was small[13]. Figure 1 illustrates the average and standard deviation of sensitivity for different types of simulated expression profiles based on dimensional reduction by k-means clustering and PCA. K-means resulted in near-zero values of sensitivity for genes with lower within-class variation of expression, such as N(5:1)|N(0;1), N(0:1)|N(5;1), N(5:1)|N(-5;1), and N(-5:1)|N(5;1). However, for PCA genes N(0;1)|N(5,1) and N(-5;1)|N(5;1) showed large negative values of sensitivity. For these genes, PCA inflated sensitivity that was not detected by k-means. One can notice in Figure 1 that, for the remainder of genes with larger differential entropy, k-means resulted in greater sensitivity when compared with PCA. Another disadvantage of PCA is that orthogonal projections may have nothing to do with class discrimination. Moreover, the bulk of data including noise and outlier patterns that often load on the lower components (>3)

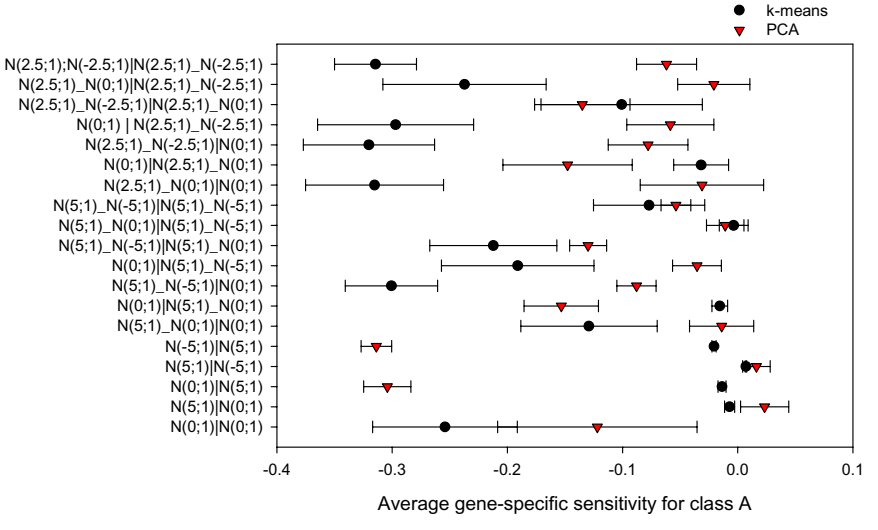

Fig. 1. Average and standard deviation of sensitivity, S_c^g, for class A

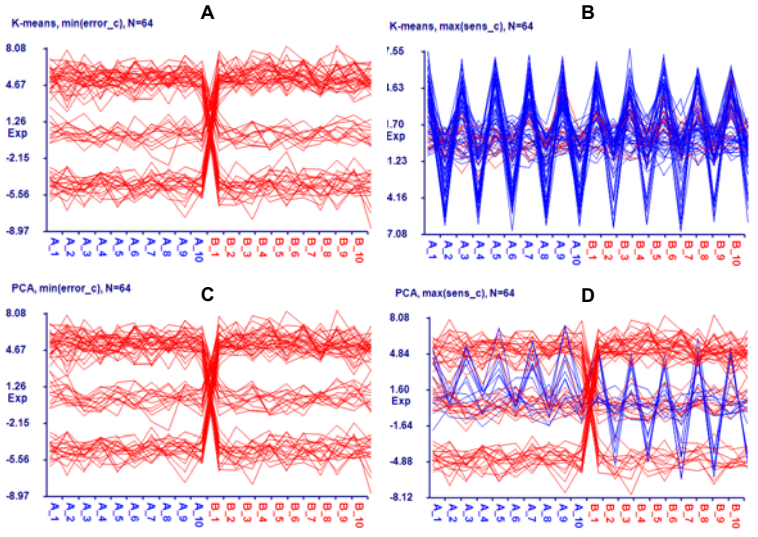

Fig. 2. Expression profiles of top 64 simulated genes identified after gene reduction with k-means or PCA followed by recursive feature elimination (RFE). (A) Dimension reduction with k-means cluster analysis and RFE based on minimum MSE. (B) K-means and maximum sensitivity. (C) PCA and minimum MSE. (D) PCA and maximum sensitivity. Red lines denote genes with significantly different ($p \leq 0.05$) expression levels between class A and B. Blue lines represent genes for which differential expression is not significantly different at the 0.05 level.

contribute less to discerning the classes. The advantage of PCA dimensional reduction is that more genes with lower within-class variation will be identified, driving up the accuracy.

Table 4 shows that, at the 0.01 level of significance, RFE based on k-means dimensional reduction along with maximum sensitivity resulted in the least amount of between-gene correlation for all 3 data sets and the least proportion of significantly differentially expressed genes for the 2-class data sets. In Figure 2B, one can visualize for simulated expression profiles that for k-means and sensitivity only 17% (0.172 from Table 4, row 2) of genes in a 64-gene model had significantly different expression. On the other hand, PCA with sensitivity (Figure 2D) resulted in 81% of the genes in a 64-gene model that had significantly different expression. The smaller proportion of significant between-gene correlation of expression due to k-means and sensitivity can also be noticed in Figure 3 for all data sets, and in particular for the Hedenfalk et al breast cancer data for which the proportion of significant ($p \leq 0.01$) between-gene correlation coefficients was 0.138.

Genes that are strongly differentially expressed and correlated may be co-regulated by shared upstream signaling molecules. A classifier based on such genes may have greater misclassification when pathway heterogeneity is important for classification. We have shown that, for the data considered, an ANN

Table 4. Recursive feature elimination (RFE) results for 64-gene ANN models including classification accuracy, proportion of significant positive between-gene correlation, and proportion of parametric tests significant among the 64 genes during original gene ranking. ANNs were fed with actual standardized expression values for the 64 genes after they were identified with RFE.

Data set	*Data reduction*	*RFE method*	*Accuracy*	$r > 0$ $p \leq 0.01$[a]	$r > 0$ $p \leq 0.05$	*Signif genes*[b]
Simulated	K-means	$\min(E_c^g)$	1.000	0.492	0.492	1.000
		$\max(S_c^g)$	0.976	0.328	0.467	0.172
	PCA	$\min(E_c^g)$	1.000	0.492	0.492	1.000
		$\max(S_c^g)$	1.000	0.341	0.371	0.813
Hedenfalk et al	K-means	$\min(E_c^g)$	1.000	0.322	0.404	0.969
		$\max(S_c^g)$	0.962	0.138	0.266	0.515
	PCA	$\min(E_c^g)$	0.967	0.197	0.325	0.703
		$\max(S_c^g)$	0.900	0.177	0.284	0.641
Khan et al	K-means	$\min(E_c^g)$	0.963	0.235	0.293	1.000
		$\max(S_c^g)$	0.960	0.200	0.270	1.000
	PCA	$\min(E_c^g)$	0.996	0.239	0.295	1.000
		$\max(S_c^g)$	0.998	0.301	0.359	1.000

[a] Proportion of 2016 between-gene correlation coefficients (i.e., $n(n-1)/2$) for 64 gene expression profiles with $p \leq 0.01$.
[b] Proportion of 64 genes with significant parametric test (t-test or F-test) during original gene ranking.

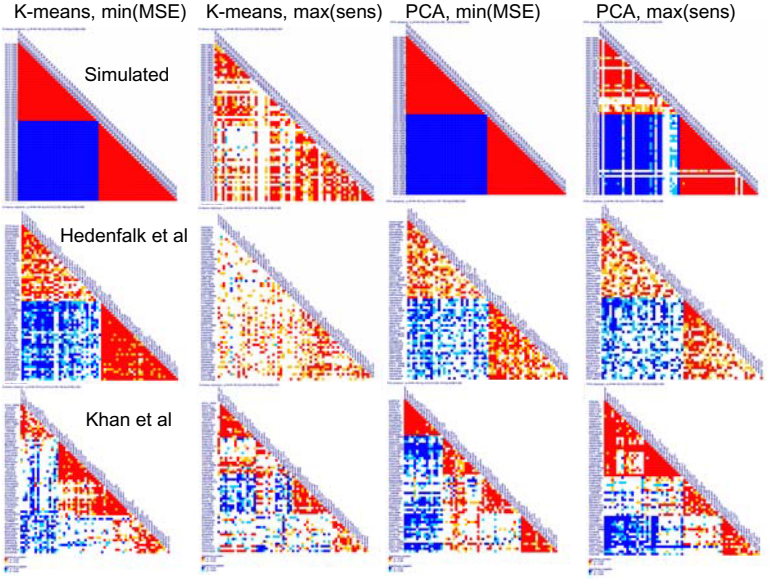

Fig. 3. Plot of significant between-gene correlation for 64-gene models. ANNs trained with standardized expression profiles for 64 genes identified during recursive feature elimination (RFE). Red denotes significant ($p \leq 0.01$) positive correlation, whereas blue signifies significant negative correlation.

classifier based on k-means dimensional reduction and sensitivity for RFE can result in accuracy levels exceeding 90% with fewer invariant and correlated features. K-means cluster analysis coupled with sensitivity for RFE may increase

detection of patients with pathway heterogeneity, which may not be tackled as well by RFE with minimum error or dimensional reduction with PCA.

References

1. Tarca, A.L., Cooke, J.E., Mackay, J.: A robust neural networks approach for spatial and intensity dependent normalization of cDNA microarray data. Bioinformatics **21** (2005) 2674-2683
2. Sawa, T., Ohno-Machado, L.: A neural network-based similarity index for clustering DNA microarray data. Comput. Biol. Med. **33** (2003) 1-15
3. Tham, C.K., Heng, C.K., Chin, W.C.: Predicting risk of coronary artery disease from DNA microarray-based genotyping using neural networks and other statistical analysis tool. J. Bioinform. Comput. Biol. **1** (2003) 521-539
4. 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. **93** (2002) 1207-1212
5. 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. **81** (2003) 594-606
6. O'Neill, M.C., Song, L.: Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. BMC Bioinformatics **4** (2003) 13
7. 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. **5** (2004) 136
8. Wei, J.S., Greer, B.T., Westermann, F., et al.: Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma. Cancer Res. **64** (2004) 6883-6891
9. Khan, J., Wei, J.S., Ringner, M., et al.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Med. **7** (2001) 673-679
10. Hedenfalk, I., Duggan, D., Chen, Y., et al.: Gene-expression profiles in hereditary breast cancer. N. Engl. J. Med. **344** (2001) 539-548
11. Hand, D.J., Mannila, H., Smyth, P.: Principles of Data Mining (Adaptive Computation and Machine Learning). Cambridge(MA): MIT Press, 2001
12. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York(NY): Springer, 2005
13. Li, T., Zhang, C., Ogihara, M.: A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics **20** (2004) 2429-2437