Derivation of Stable Microarray Cancer-Differentiating Signatures Using Consensus Scoring of Multiple Random Sampling and Gene-Ranking Consistency Evaluation

Zhi Qun Tang,[1,2] Lian Yi Han,[1,2] Hong Huang Lin,[1,2] Juan Cui,[1,2] Jia Jia,[1,2] Boon Chuan Low,[2,3] Bao Wen Li,[2,4] and Yu Zong Chen[1,2]

[1]Bioinformatics and Drug Design Group, Department of Pharmacy; [2]Center for Computational Science and Engineering; and Departments of [3]Biological Sciences and [4]Physics, National University of Singapore, Singapore, Singapore

Abstract

Microarrays have been explored for deriving molecular signatures to determine disease outcomes, mechanisms, targets, and treatment strategies. Although exhibiting good predictive performance, some derived signatures are unstable due to noises arising from measurement variability and biological differences. Improvements in measurement, annotation, and signature selection methods have been proposed. We explored a new signature selection method that incorporates consensus scoring of multiple random sampling and multistep evaluation of gene-ranking consistency for maximally avoiding erroneous elimination of predictor genes. This method was tested by using a well-studied 62-sample colon cancer data set and two other cancer data sets (86-sample lung adenocarcinoma and 60-sample hepatocellular carcinoma). For the colon cancer data set, the derived signatures of 20 sampling sets, composed of 10,000 training test sets, are fairly stable with 80% of top 50 and 69% to 93% of all predictor genes shared by all 20 signatures. These shared predictor genes include 48 cancer-related and 16 cancer-implicated genes, as well as 50% of the previously derived predictor genes. The derived signatures outperform all previously derived signatures in predicting colon cancer outcomes from an independent data set collected from the Stanford Microarray Database. Our method showed similar performance for the other two data sets, suggesting its usefulness in deriving stable signatures for biomarker and target discovery. [Cancer Res 2007;67(20):9996–10003]

Introduction

Microarrays have been explored for deriving molecular signatures, which are subsets of genes differentially expressed in patients of different disease outcomes, for disease diagnosis and prognosis (1–6), and for determining disease mechanisms (1, 7), targets (8), and treatment strategies (9, 10). Although showing good predictive performance (4, 5, 9, 11–13), the derived signatures have been found to be highly unstable and to include fewer disease-related genes (9, 14, 15). For instance, 10 different sets of signatures for separating colon cancer tissues and normal colon tissues have been derived from the same 62-sample data set (16) by using different sampling methods. Only 1 to 5 of the 4 to 60 selected predictor genes in each of these sets are present in more than half of the other nine sets (Table 1), and 2 to 20 of the predictor genes in each set are cancer related (Table 2). Despite the use of sophisticated class differentiation and signature selection methods, the selected signatures show few overlapping predictor genes, as in the case of other microarray data sets including non–Hodgkin lymphoma, acute lymphocytic leukemia, breast cancer, lung adenocarcinoma, medulloblastoma, hepatocellular carcinoma, and acute myeloid leukemia (9, 15).

Although these signatures display high cancer differentiation accuracies at levels of 85.9% to 100%, the highly unstable and patient-dependent nature of these signatures diminishes their application potential for diagnosis and prognosis (9). Moreover, the complex and heterogenic nature of cancer may not be adequately described by the few cancer-related genes in some of these signatures. The unstable nature of these signatures and their lack of disease-relevant genes also limit their potential for target discovery. The instability of derived signatures is likely caused by the noises in the microarray data arising from measurement variability and biological differences (9, 14). These noises arise from such factors as the precision of measured absolute expression levels, capability for detecting low abundance genes, quality of design and probes, annotation accuracy and coverage, and biological differences of expression profiles (14, 17). Apart from enhancing the quality of measurement and annotation (17), strategies for improving signature selection have also been proposed. These include the use of multiple random validation (9), larger sample size (18), known mechanisms (19), and signature selection methods less sensitive to noises (1, 14, 20).

We explored a new signature selection method aimed at reducing the chances of erroneous elimination of predictor genes due to the noises contained in microarray data set. In our approach, non–predictor genes were eliminated by consensus scoring of a large number of training and test sets generated from repeated random sampling (9), and by incorporating a multistep gene-ranking consistency evaluation procedure into a well-established signature selection method. Our method was tested by using a well-studied colon cancer data set (16) and two other data sets—86 samples of lung adenocarcinoma (21) and 60 samples of hepatocellular carcinoma (22) so that our derived signatures can be adequately evaluated and compared with those of other studies using different sampling sets, class differentiation methods, and feature selection methods. The performance of selected signatures for the colon cancer data set was further evaluated by applying the selected signatures in predicting colon cancer outcomes from an independent colon cancer data set separately generated from Stanford Microarray Database (23), and the

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Requests for reprints: Yu Zong Chen, Department of Pharmacy, National University of Singapore, S16, Level 8, 6 Science Drive 2, Singapore 117546, Singapore. Phone: 65-6516-6877; Fax: 65-6774-6756; E-mail: phacyz@nus.edu.sg.

doi:10.1158/0008-5472.CAN-07-1601

performance of these signatures was compared with those of previously derived colon cancer signatures applied to the same data set.

Materials and Methods

Classification method. Support vector machines (SVM), a supervised machine learning method, was used for training a class differentiation system (14, 24, 25). In classification of microarray data sets, it has been found that supervised machine learning methods generally yield better results (26), particularly for smaller sample sizes (14). In particular, SVM consistently shows outstanding performance, is less penalized by sample redundancy, and has lower risk for overfitting (25, 27).

SVMs (24) project feature vectors into a high-dimensional feature space by using a kernel function $k(X, X') = \exp\left(-\frac{\|X-X'\|^2}{2\sigma^2}\right)$. The linear SVM procedure can then be applied to the feature vectors in this feature space: It constructs a hyperplane that separates two different classes of feature vectors with a maximum margin. This hyperplane is constructed by finding a vector $\mathbf{w}$ and a variable b that minimizes $\|\mathbf{w}\|^2$, which satisfies the following conditions: $\mathbf{w}\cdot\mathbf{x}_i + b \geq +1$, for $y_i = +1$ (cancer patients) and $\mathbf{w}\cdot\mathbf{x}_i + b \leq -1$, $y_i = -1$ (normal people). Here, $\mathbf{x}_i$ is a feature vector, y_i is the group index, $\mathbf{w}$ is a vector normal to the hyperplane, $|b| / \|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin, and $\|\mathbf{w}\|$ is the Euclidean norm of $\mathbf{w}$. After the determination of $\mathbf{w}$ and b, a given vector $\mathbf{x}$ can be classified by using sign $[(\mathbf{w}\cdot\mathbf{x}) + b]$; a positive or negative value indicates that the vector $\mathbf{x}$ belongs to the positive or negative class, respectively.

The performance of SVM classification can be measured by true positive *TP* (number of cancer patients correctly predicted as cancer patient), false negative *FN* (number of cancer patients incorrectly predicted as normal), true negative *TN* (number of normal correctly predicted normal), and false positive *FP* (number of normal incorrectly predicted as cancer patients). Three indicators, sensitivity $Q_p = TP / (TP + FN)$, specificity $Q_n = TN / (TN + FP)$, and overall accuracy $Q = (TP + TN) / (TP + FN + TN + FP)$, were used to measure the predictive performance.

Feature selection method. Predictor genes of each training test set were selected by using SVM recursive feature elimination (RFE-SVM), which is a wrapper method that selects predictor genes by eliminating non–predictor genes according to a gene-ranking function generated from a class differentiation system (28). Wrapper methods generally perform better than other feature selection methods (28). RFE-SVM is the best performing wrapper method and has thus been more widely used in cancer microarray analysis (24, 27).

The ranking criterion of RFE-SVM is based on the change in the objective function upon removing each feature. To improve the efficiency of training, this objective function is represented by a cost function J for the kth feature computed by using training set only. When a given feature is removed or its weight w_k is reduced to zero, the change in the cost function $J(k)$ is given by $DJ(k) = \frac{1}{2}\frac{\partial^2 J}{\partial w_k^2}(Dw_k)^2$. The case of $Dw_k = w_k - 0$ corresponds to the removal of feature k. In our case, the change in the cost function can be estimated as $DJ(k) = \frac{1}{2}\alpha^T H\alpha - \frac{1}{2}\alpha^T H(-k)\alpha$, where H is the matrix with elements $y_i y_j exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$. $H(-k)$ is the matrix computed by using the same method as that of matrix H but with its kth component removed. The change in the cost function indicates the

Table 1. Statistics of the selected molecular signatures for differentiating microarray data set of colon cancer patients from that of normal people by 10 different studies that used the same data set

Study (reference)	No. selected genes in signature	Class differentiation method	Signature selection method	Validation method	Prediction accuracy	No. genes selected by other *n* studies									
						9	8	7	6	5	4	3	2	1	0
Zhou and Mao, 2005 (29)	15	LS-SVM	A hybrid of filter and wrapper methods (LS bound measure)	Bootstrap	<85%	0	0	0	1	0	1		1	1	11
Ding and Peng, 2005 (30)	60	NB, SVM, LDA, LR	Filter method (MRMR)	LOOCV	93.55%	0	0	0	3	2	1	1	4	4	45
Guyon et al., 2002 (24)	7	SVM (linear kernel)	Wrapper method (RFE)	LOOCV	98%	0	0	0	0	0	1	1	2	0	3
Inza et al., 2004 (28)	5	Decision tree 1	Wrapper method	LOOCV	87.1%	0	0	0	3	2	0	0	0	0	0
Inza et al., 2004 (28)	4	Decision tree 2	Wrapper method	LOOCV	88.81%	0	0	0	0	0	1	1	2	0	0
Bo and Jonassen, 2002 (31)	50	Linear discriminant	Gene pair ranking	LOOCV	87.8%	0	0	0	3	2	1	2	7	8	27
				L-31-OCV	85.9%										
Huang and Kecman, 2005 (32)	10	SVM1	Wrapper method (RFE)	LOOCV	Not indicated	0	0	0	3	2	0	0	5	0	0
Huang and Kecman, 2005 (32)	10	SVM2	Wrapper method (RFE)	LOOCV	88.84%	0	0	0	3	0	0	0	2	2	3
Huang and Kecman, 2005 (32)	10	SVM3	Wrapper method (RFE)	LOOCV	88.1%	0	0	0	3	2	0	0	4	0	1
Liu et al., 2005 (33)	6	Clustering method	Filter method (mutual information)	LOOCV	91.9%	0	0	0	2	2	0	0	0	2	0
Total no. uniquely selected genes, 107				No. unique genes selected by only one study, 83											

NOTE: The data set is from ref. 16. The authors of the second study tried four methods (naïve Bayes classifier, SVM, linear discriminant analysis, and logistic regression) and choose one that showed the highest accuracy.
Abbreviations: LS-SVM, least square SVM; MRMR, minimum redundancy-maximum relevance feature selection framework; NB, naïve Bayes classifier; LDA, linear discriminant analysis; LR, logistic regression; LOOCV, leave-one-out cross-validation; SVM1, SVM2, and SVM3, SVM classifiers with different variables; decision tree 1, decision tree 2, decision tree classifiers with different variables.

contribution of the feature to the decision function and serves as an indicator of gene ranking position.

Gene-ranking consistency evaluation. The microarray data set was randomly divided into a training set (contains half of the samples) and an associated test set (the other half of the samples). By using repeated random sampling (9), 10,000 training test sets, each containing a unique combination of samples, were generated. These 10,000 training test sets were randomly placed into 20 sampling groups; each group contains 500 training test sets. Every sampling group was then used to derive a signature based on consensus scoring and evaluation of gene-ranking consistency of the corresponding 500 training and 500 test sets. The 20 different signatures derived from these sampling groups were compared to test the level of stability of selected predictor genes.

In each group containing 500 training and 500 associated test sets, gene subsets were selected by RFE-SVM from each training set and the performance of gene subsets were evaluated from the associated test set. To derive a gene ranking criterion consistent for all iterations, RFE gene ranking function at every iteration step was derived from a SVM class differentiation system with a universal set of globally optimized variables that gave the best average class differentiation accuracy over the 500 test sets.

To reduce the chance of erroneous elimination of predictor genes due to noises in microarray data, additional gene-ranking consistency evaluation steps were implemented on top of the normal RFE procedures in all sampling sets. In step 1, for every test set, subsets of genes ranked in the bottom 10% (if no gene was selected in current iteration, this percentage was gradually increased to the bottom 40%) with combined score lower than the first few top-ranked genes were selected such that the collective contribution of these genes will less likely outweigh higher-ranked ones. In step 2, for every test set, the step 1 selected genes was further evaluated to choose those not ranked in the upper 50% in previous iteration so as to ensure that these genes are consistently ranked lower. In step 3, a consensus scoring scheme was applied to step 2 selected genes such that only those appearing in >90% (if no gene was selected in current iteration, this percentage was gradually reduced to 60%) of the 500 test sets were eliminated.

For each sampling set, different SVM variables were scanned; various RFE iteration steps were evaluated to identify the globally optimal SVM variables and RFE iteration steps that give the highest average class differentiation accuracy for the 500 test sets. The 20 different signatures derived from these sampling sets were then compared to test the level of stability of selected predictor genes.

Results and Discussion

Colon cancer data set. The 62-sample colon cancer data set contains the measurements for 2,000 gene probes, of which 40 samples were labeled as tumor and 22 as normal (16). Data were subjected to the standard preprocessing procedure (24). These data have been analyzed in several previous studies using SVM or other statistical approaches (24, 28–33). In multiple random sampling, this data set was randomly divided into a training set containing 31 samples and an associated test set containing the other 31 samples. Twenty signatures were derived from 10,000 training test sets that were generated as described before.

The stability levels of the 20 derived signatures can be estimated from the percentage of predictor genes shared by all 20 signatures. From Table 3, 80% of the top 50 ranked genes and 69% to 93% of all genes in each signature were shared by 20

Table 2. Distribution of the selected predictor genes of the 10 studies in Table 1 with respect to different cancer-related classes

Study (reference)	Cancer genes						Tumor markers	Interacting partner of cancer genes	Cancer pathway affiliated genes	Genes having possible implication in any cancer
	Anticancer targets	Oncogenes	Tumor-suppressor genes	Angiogenesis genes	Immune tolerance genes	Other types				
Zhou and Mao, 2005 (LS-SVM; ref. 29)	2	0	0	0	0	5	0	0	0	0
Ding and Peng, 2005 (NB, SVM, LDA, LR; ref. 30)	1	0	1	0	0	5	1	4	2	1
Guyon et al., 2002 (SVM; ref. 24)	1	0	0	0	0	0	0	0	1	0
Inza et al. 2004 (decision tree 1; ref. 28)	0	0	0	0	0	1	0	0	1	0
Inza et al., 2004 (decision tree 2; ref. 28)	0	0	0	0	0	0	0	0	0	0
Bo and Jonassen, 2002 (linear discriminant; ref. 31)	2	2	2	0	0	4	0	7	3	0
Huang and Kecman, 2005 (SVM1; ref. 32)	0	0	0	0	0	2	0	2	1	0
Huang and Kecman, 2005 (SVM2; ref. 32)	0	0	2	0	0	1	0	1	2	0
Huang and Kecman, 2005 (SVM3; ref. 32)	0	0	0	0	0	1	0	2	1	0
Liu et al., 2005 (clustering method; ref. 33)	1	0	0	0	0	0	0	1	1	0

Table 3. Statistics of the selected predictor genes for predicting cancer outcome from a colon cancer data set by class differentiation systems constructed from 20 different sampling sets each composed of 500 training test sets generated by random sampling

Sampling set	No. selected predictor genes in signature	No. predictor genes also included in *n* other signatures derived by using different sampling set																			
		19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
1	155	104	3	2	5	3	4	3	4	3	3	1	4	2	1	4	2	3	3	1	0
2	135	104	3	2	4	3	4	2	4	2	1	0	1	2	1	0	1	1	0	0	0
3	156	104	3	2	5	3	4	3	3	3	3	1	3	3	0	4	4	4	0	1	3
4	146	104	3	2	5	3	4	3	3	3	2	1	2	2	1	2	2	1	2	0	1
5	116	104	3	1	4	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0
6	112	104	2	0	2	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0
7	119	104	2	2	3	1	3	0	3	0	0	1	0	0	0	0	0	0	0	0	0
8	127	104	3	2	5	3	2	1	2	1	1	1	1	0	0	1	0	0	0	0	0
9	133	104	3	2	5	2	2	2	2	2	2	0	2	1	2	1	0	0	0	0	1
10	156	104	3	2	4	3	4	3	4	3	3	0	3	2	1	2	3	5	3	2	2
11	139	104	3	2	5	3	3	3	3	3	2	1	2	1	1	3	0	0	0	0	0
12	115	104	2	2	3	1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0
13	144	104	3	2	5	3	4	3	3	2	3	1	2	2	1	3	0	1	2	0	0
14	157	104	3	2	5	3	4	3	4	3	3	0	3	2	2	3	2	3	3	3	2
15	149	104	3	2	4	3	4	3	3	3	3	0	3	2	0	3	1	3	2	0	3
16	136	104	3	2	5	3	4	3	3	2	0	0	3	2	1	0	0	0	1	0	0
17	136	104	3	2	5	3	3	2	3	1	3	0	2	1	1	0	1	1	1	0	0
18	127	104	3	2	4	3	2	3	1	0	1	0	2	0	0	1	1	0	0	0	0
19	146	104	3	2	5	2	4	3	4	3	2	1	2	2	1	1	3	1	1	1	1
20	122	104	3	1	2	2	2	1	1	2	1	1	0	0	1	1	0	0	0	0	0

signatures. This suggests that our selected signatures are fairly stable. One reason is that a SVM class differentiation system with a universal set of globally optimized variables, which gave the best average class differentiation accuracy over the 500 test sets, was used to derive RFE gene ranking function at every iteration step and for every test set. In earlier studies using RFE or other wrapper methods for selecting signatures, non–predictor genes have been eliminated in multiple iterations, and at every iteration step a different class differentiation system, characterized by a different set of optimized variables, has been constructed (24, 27). As gene elimination is variable dependent, these selected predictor genes are likely path dependent and heavily influenced by sampling method, composition, order of gene evaluation, computational algorithm, and variables. These characteristics partly explain the highly unstable and patient-dependent characteristics of the previously derived signatures (16). Another reason is that an additional gene-ranking consistency evaluation is done on top of the normal RFE procedure to reduce the change of erroneous elimination of predictor genes.

The optimal SVM variables for the 20 sample sets are in a narrow range of 17 to 18 and the highest average accuracies were 92.2% to 92.8% for tumor patients and 90.4% to 91.1% for normal people, respectively. At these variables, the accuracies for the individual test sets ranged from 82.4% to 100% for tumor patients and 77.0% to 100% for normal people, respectively. Further deviation from these optimal variables has relatively smaller effect on prediction accuracy and composition of predictor genes. The relatively small variations of optimal SVM variables and prediction accuracies across the 20 sampling sets suggest that the performance of the SVM class differentiation systems constructed by using globally optimized variables and RFE iteration steps are fairly stable across different sampling combinations.

The number of predictor genes in our signatures ranges from 112 to 157 (Table 3), which is substantially higher than those of 6 to 60 in the previously derived signatures. It was reported that there are 291 known cancer genes (34), 15 cancer-associated pathways (35), 34 angiogenesis genes (36, 37), and 43 tumor immune tolerance genes (38). Because of biological differences and complex nature of cancers, a signature applicable for many patients is expected to include a substantial percentage of these cancer-related genes, together with some of their interacting partners and consequence genes (34). Moreover, because of measurement variability, a certain number of irrelevant genes may be inevitably included in a signature. Therefore, it is not surprising to find cancer signatures with 110 to 150 predictor genes. Moreover, for target discovery, which is a very important application of gene selection from microarray analysis, it is probably unrealistic to assume that only a few genes stand out from the thousands of gene with sufficient clarity to allow target selection (8).

The 104 predictor genes shared by all 20 signatures (Table 4; Supplementary Table S1 and S2) include 48 cancer-related genes (4 anticancer targets, 3 oncogenes, 8 tumor suppressors, 2 angiogenesis genes, 1 tumor immune tolerance gene, 4 cancer genes, 3 tumor markers, 17 cancer-gene interacting genes, and 6 cancer pathway–affiliated genes). In our analysis, anticancer targets were obtained from the latest version of therapeutic target database[5] (39, 40), and the cancer-related genes and cancer pathways were taken from recent publications (34–38, 41) and references in Supplementary Table S1.

[5] http://bidd.nus.edu.sg/group/cjttd/ttd.asp

These 104 shared predictor genes also include 16 genes possibly implicated in cancer (description and references in Supplementary Table S1). They have been reported to be involved in cancer risk (*GSTM4*), promotion of metastatic growth (*POSTN*) and tumor invasion (*SERPINE2*), maintenance of telomere repeats in cancer cells (*HNRPA1*), energy metabolism in cancer cells (*PCCB*), regulation of cell cycle (*CSRP1, EEF1B2, TSPAN1, WDR7*) and gene expression (*CNOT1*), and synthesis of signaling molecules with elevated levels in tumors (*GUCA2B*). Genes reported to have significantly altered expression level in tumors with unclear connection to cancer (*PMP22, SRPK1, UQCRC1*) were also included here because of their possible roles as cancer consequence genes.

With a significantly higher number of cancer-related genes than those of 2 to 20 in the 10 previously derived signatures, our signatures seem to more closely reflect the complex nature of cancer known to involve collective actions of many genes of different functions (34–38, 41). Moreover, our signatures include 52 of the 107 previously derived predictor genes, and those selected by a higher number of other studies tend to be ranked higher by our gene-ranking function (Supplementary Table S2). Regardless of their possible roles in cancer, these genes have shown proven capability for colon cancer outcome prediction. It is not surprising that they are included in our signatures.

To further evaluate the predictive capability of our selected signatures, we collected the gene expression profiles of 34 colon cancer cell lines and 8 normal colon tissues from the Stanford Microarray Database (23). The predictive capability of our selected and the 10 other previously derived signatures were evaluated by using the SVM classification system and 500 randomly generated training test sets generated from this data set using the same procedure described in Materials and Methods. The performance was evaluated by using the associated test set, which is shown in Table 5. The overall accuracy for the 104 predictor genes is 96.8%, with a SD of 3.3%. Using genes selected by other methods, the overall accuracies were found to range from 80.5% to 94.9%, with a SD of 2.9% to 6.6%. These results suggest that the signatures selected using our method can perform more stable and better than those selected by other methods.

The predictive capability of our selected and the 10 other previously derived signatures were also evaluated by using additional 500 randomly generated training test sets generated from the original colon cancer microarray data set and contained different combination of samples from those used for gene signatures selection. The average cancer differentiating accuracies of our signatures over these 500 test sets are, respectively, 92.0% to 92.1%, 91.8% to 92.0%, 90.7% to 91.4%, and 89.1% to 89.8% when all, top 100, top 50, and top 30, predictor genes were used. The SDs of the individual accuracies are in the range of 3.4% to 3.5%, 3.4% to 3.6%, 3.7% to 3.9%, and 6.4% to 6.7%. In contrast, the average accuracies and SDs of the 10 previously derived signatures are in the range of 63.8% to 89.9% and 4.6% to 12.3%. Hence, the performance of our signatures is more stable than those of other signatures. The top 50 predictor genes perform almost as well as the full features, whereas the performance of the top 30 predictor genes is substantially less stable.

It has been reported that six samples in the colon cancer data set might have been wrongly labeled (42). These include three tumor tissues (T33, T36, T30) more probable to be normal ones and three normal tissues (N8, N34, N36) more likely to be cancerous. Four of these six samples (T33, T36, T30, N36) are misclassified as their opposite labels by >90% of the 500 SVM models. Another one (N34) is misclassified in 74% of the 500 SVM models. Misclassification of T33, T36, and T30 into their opposite labels is actually consistent with the opinion that these are more likely normal tissues.

Table 4. List of predictor genes of colon cancer data set shared by all 20 signatures

Gene group	Predictor genes selected by both this work and other studies (no. studies)	Predictor genes selected by this work only
Anticancer targets (successful or research)	*CDKN1A* (2), *HSD11B2*(1)	*MMP9, SPARC*
Cancer genes (oncogenes)	*HMGA1* (1), *ETS2* (1)	*NCOA4*
Cancer genes (tumor suppressors)	*MXI1* (1), *TPM1* (2), *CDH3* (2)	*CDX1, PRKACA, CREB5, PPP2R5C, CNNM4*
Cancer genes (angiogenesis genes)		*PECAM1, CCL14*
Tumor immune tolerance genes		*IL1R2*
Cancer genes (other types)	*MYH9* (7), *S100P* (3), *HSP90AB1* (1)	*IFI6*
Tumor markers	*MUC2* (1)	*S100A11, KRT8*
Interacting partner of cancer genes	*TPM2* (3), *CKS2* (2), *EIF2S2* (2), *HSPD1* (2), *MT1G* (2), *GSN* (1), *MORF4L2* (1), *SND1* (1), *TCF8* (1)	*ATP2A2, CD46, CPSF1, PTPRH, WASF2, PPP1R9B, CALM2, CBX3*
Cancer pathway affiliated genes	*DES* (7), *C15orf15* (3), *MYL6* (2)	*RPS6, RPL30, POLD2*
Genes having possible implication in cancer	*GUCA2B* (6), *CSRP1* (5), *SPARCL1* (3), *HNRPA1* (1), *SERPINE2* (1), *TSPAN1* (1)	*CNOT1, EEF1B2, PCCB, PMP22, POSTN, SRPK1, WDR7, GSTM4, UQCRFS1, UQCRC1*
Others	*MYL9* (7), *WDR77* (4), *DARS* (3), *VIP* (3), *ATP5J* (3), *GTF3A* (1), *NR3C2, MLC1* (1), *GPD1L* (1), *CFD* (1), *ZNF358* (1), *FUCA1* (1), *GABRB3* (1), *SNRPB* (1), *DTWD2* (1), *CST3* (1), *FBL* (1), *MARCKSL1* (1), *SCNN1B* (1), *H11084* (1), *H40095* (1), *H64807* (1)	*ACAA2, ARL6IP, CLNS1A, COX6A1, COX8A, FXYD1, GYG1, IFITM2, ITPR3, KIF5A, MGC22793, MT1M, PCNP, PRPSAP1, H73908*

NOTE: Alternative names and function of these genes are provided in Supplementary Table S1.

Table 5. Average cancer outcome prediction accuracy and SD of 500 SVM class differentiation systems constructed by 42 samples collected from Stanford Microarray Database (23) and by using signatures derived from this work and 10 previous works

Signature (method)	No. selected predictor genes in signature	Average accuracy (%)	SD (%)
All predictor genes in signature (this work)	112–157	95.8–97.4	4.7–5.0
One hundred four genes selected by all signatures (this work)	104	96.8	3.3
Ding and Peng, 2005 (NB, SVM, LDA, LR; ref. 30)	60	94.9	5.9
Huang and Kecman, 2005 (SVM3; ref. 32)	10	94.8	6.5
Bo and Jonassen, 2002 (linear discriminant; ref. 31)	50	94.6	6.5
Huang and Kecman, 2005 (SVM2; ref. 32)	10	93.7	6.5
Huang and Kecman 2005 (SVM1; ref. 32)	10	93.2	6.6
Liu et al., 2005 (clustering method; ref. 33)	6	92.8	6.7
Zhou and Mao, 2005 (LS-SVM; ref. 29)	15	88.9	6.0
Inza et al., 2004 (decision tree 2; ref. 28)	4	86.0	5.4
Inza et al., 2004 (decision tree 1; ref. 28)	5	82.9	4.3
Guyon et al., 2002 (SVM; ref. 24)	7	80.5	2.9

Likewise, misclassification of N36 and N34 is consistent with the opinion that they are more likely cancerous. Despite of the "incorrect" labeling of six samples, our SVM models are "fooled" by only one of these samples. These results suggest that our method and derived SVM models are less sensitive to incorrect labeling of a small percentage of samples.

Lung adenocarcinoma and hepatocellular carcinoma data set. The lung adenocarcinoma data set contains the expression profiles of 7,129 genes from 86 lung adenocarcinoma patients (21). These 86 patients have been divided into two groups, survivable (62 patients) and nonsurvivable (24 patients), based on whether the patient was still alive in a postsurgery follow-up survey (21). This data set[6] has been analyzed in several previous studies (9, 43, 44). Although these studies show good performances for separating the survivable and nonsurvivable patients, few of the selected predictor genes are shared by these reported signatures. In this work, the relevant data was subjected to the standard preprocessing procedure, as described by Guyon et al. (24).

In multiple random sampling, this data set was randomly divided into a training set containing 43 samples and an associated test set containing the other 43 samples. To reduce computational cost, 3,000 training test sets, each containing a unique combination of samples, were generated. These 3,000 training test sets were randomly placed into six sampling groups, each containing 500 training test sets. Every sampling group was then used to derive a signature based on consensus scoring and evaluation of gene-ranking consistency of the corresponding 500 training and 500 test sets. Finally, the six different signatures derived from these sampling groups were compared with test the level of stability of selected predictor genes.

The results are summarized in Table 6. The number or predictor genes in our signatures ranges from 42 to 56. A total of 36 predictor genes, representing 64.3% to 85.7% of all genes in each signature, were shared by all six signatures. The predictive capability of our selected signatures was evaluated by using an additional 500 randomly generated training test sets generated from the original lung adenocarcinoma microarray data set. The average survival prediction accuracies of our signatures over these 500 test sets are 95.5% to 96.7%, which are comparable with those derived by other studies but our signatures are significantly more stable as manifested by the high percentages of selected predictor genes shared by all signatures. These results suggest that our method is capable of selecting fairly stable signatures for predicting lung adenocarcinoma survivability.

Based on our selected 36 predictor genes, the 86 lung adenocarcinoma patients were grouped into two clusters by using unsupervised hierarchical clustering method (Supplementary Fig. S1*A*). Kaplan-Meier survival analysis has shown that the survival time in these two patient clusters was significantly different ($P < 0.0001$, log-rank test, Supplementary Fig. S1*B*). Cluster 1 is the poor prognosis group with 45.1 months of average survival time. Cluster 2 is the good prognosis group with 96.6 months of average survival time. This suggests that our predictor genes are useful for differentiating different survival groups by using a different method.

The hepatocellular carcinoma oligonucleotide array data set (22) contains the expression profiles of 7,129 genes from 60 hepatocellular carcinoma patients. These 60 patients have been divided into two groups, recurrent (20 patients) and nonrecurrent (40 patients), based on a postsurgery follow-up survey (22). This data set[7] has been analyzed in several previous works (9, 45). Although these studies show good performances for separating the recurrent and nonrecurrent patients, few of the selected predictor genes are shared by these reported signatures. The relevant data was also processed by using the same standard preprocessing procedure as that described in the literature (24).

Using the same procedure as that of the lung adenocarcinoma data set, six different sets of hepatocellular carcinoma recurrence signatures were derived from 3,000 training test sets. The results are summarized in Table 6. The number of predictor genes in our signatures ranges from 25 to 30 of which 16 genes were shared by

[6] http://dot.ped.med.umich.edu:2000/ourimage/pub/Lung/index.html

[7] http://surgery2.med.yamaguchi-u.ac.jp/research/DNAchip/hcc-recurrence/index.html

Table 6. The summary of signature selection results from lung adenocarcinoma and hepatocellular carcinoma microarray data sets

Microarray data set	Lung adenocarcinoma (21)	Hepatocellular carcinoma (22)
Clinical end point	Survival	Early recurrence
Sample size	86	60
No. signatures	6	6
No. selected predictor genes in each signature (number of genes shared by all signatures)	42–56 (36)	25–30 (16)
Percentage of predictor genes shared by all signatures (%)	64.3–85.7	53.3–64.0
Average prediction accuracy	95.5–96.7	98.1–99.3
SD	3.2–3.7	2.9–3.5
Selected genes	*PRKACB, LDHB, VEGF, PLD1, FXYD3, ADFP, SPRR1B, GRO3, RPS3, GALNT4, POLD3, FUT3, SLC2A1, BSG, REG1A, SERPINE1, VDR, TUBA1, ANXA8, RDC1, CHRNA2, NULL(HG2175-HT2245), GARS, CYP24, KRT18, HLA-G, SPRR2A, CD58, FEZ2, WNT10B, E48, ALDH2, STX1A, KRT14, ENO2, SCYB5*	*SRBP1, PPARP1, APB21, HAX-1, NDP52, IL1RA, KIAA0235, MBL2, IFNRTF, FADD, G3BP1, PRAD1, SGK1, SDF1b, HG2724-HT2820, CLCKB*

NOTE: The prediction accuracy was obtained from 500 SVM class differentiation systems constructed by using each of the signatures derived from this work.

all the six signatures. This indicated that 53.3% to 64.0% of all genes in each signatures were shared by six signatures. The predictive capability of our selected signatures were evaluated by using the SVM classification system and 500 randomly generated training test sets generated from this hepatocellular carcinoma data set using the same procedure described in Materials and Methods. The performance was evaluated by using the associated test set and the average accuracies of our signatures over these 500 test sets were found to range from 98.1% to 99.3%, which is comparable with those derived by other studies (9, 45) but our signatures are significantly more stable as they contain a high percentage of shared predictor genes.

It is noted that the numbers of predictor genes in the lung adenocarcinoma and hepatocellular carcinoma data sets are significantly less than the number of predictor genes from the colon cancer data set. One possible reason for this difference is that the expression profiles of some cancer genes important for differentiating cancer and noncancer patients may not be significantly different in cancer patients of different survival groups or recurrent groups. As a result, higher number of cancer genes is expected to be selected in the signatures of the colon cancer data set than those of the lung adenocarcinoma and hepatocellular carcinoma data sets.

Summary

Our study suggests that fairly stable signatures can be selected from microarray data sets by using our proposed new feature selection method. The use of consensus scoring for multiple random sampling (9) and evaluation of gene-ranking consistency seem to have impressive capability in avoiding erroneous elimination of predictor genes due to such noise such as measurement variability and biological differences. Further improvement in measurement quality, annotation accuracy and coverage, and signature selection algorithms (9, 17–20) will enable the derivation of more accurate signatures for facilitating biomarker and target discovery.

Acknowledgments

Received 5/2/2007; accepted 8/3/2007.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

References

1. Winegarden N. Microarrays in cancer: moving from hype to clinical reality. Lancet 2003;362:1428.
2. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. Nat Genet 2003;33:49–54.
3. Staudt LM. Molecular diagnosis of the hematological cancers. N Engl J Med 2003;348:1777–85.
4. Bullinger L, Dohner K, Bair E, et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. N Engl J Med 2004;350:1605–16.
5. Valk PJM, Verhaak RGW, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med 2004;350:1617–28.
6. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–11.
7. Garber K. Genomic medicine. Gene expression tests foretell breast cancer's future. Science 2004;303:1754–5.
8. Meltzer PS. Spotting the target: microarrays for disease gene discovery. Curr Opin Genet Dev 2001;11:258–63.
9. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005;365:488–92.
10. Caldas C, Aparicio SA. The molecular outlook. Nature 2002;415:484–5.
11. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
12. 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.
13. Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet 2003;362:1439–44.

14. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 2006;7:55–65.
15. Bullinger L, Valk PJ. Gene expression profiling in acute myeloid leukemia. J Clin Oncol 2005;23:6296–305.
16. Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A 1999;96:6745–50.
17. Draghici S, Khatri P, Eklund AC, Szallasi Z. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet 2006;22:101–9.
18. Ioannidis JP. Microarrays and molecular research: noise discovery? Lancet 2005;365:454–5.
19. Gardner SN, Fernandes M. Prediction of cancer outcome with microarrays. Lancet 2005;365:1685.
20. Biganzoli E, Lama N, Ambrogi F, Antolini L, Boracchi P. Prediction of cancer outcome with microarrays. Lancet 2005;365:1683.
21. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816–24.
22. Iizuka N, Oka M, Yamada-Okabe H, et al. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. Lancet 2003;361:923–9.
23. Gollub J, Ball CA, Binkley G, et al. The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res 2003;31:94–6.
24. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002;46:389–422.
25. Pochet N, De Smet F, Suykens JA, De Moor BL. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. Bioinformatics 2004;20:3185–95.
26. Qiu P, Wang ZJ, Liu KJ. Ensemble dependence model for classification and prediction of cancer and normal gene expression data. Bioinformatics 2005;21:3114–21.
27. Li F, Yang Y. Analysis of recursive gene selection approaches from microarray data. Bioinformatics 2005;21:3741–7.
28. Inza I, Larranaga P, Blanco R, Cerrolaza AJ. Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med 2004;31:91–103.
29. Zhou X, Mao KZ. LS Bound based gene selection for DNA microarray data. Bioinformatics 2005;21:1559–64.
30. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 2005;3:185–205.
31. Bo T, Jonassen I. New feature subset selection procedures for classification of expression profiles. Genome Biol 2002;3:RESEARCH0017.
32. Huang TM, Kecman V. Gene extraction for cancer diagnosis by support vector machines—an improvement. Artif Intell Med 2005;35:185–94.
33. Liu X, Krishnan A, Mondry A. An entropy-based gene selection method for cancer classification using microarray data. BMC Bioinformatics 2005;6:76.
34. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177–83.
35. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004;10:789–99.
36. de Castro Junior G, Puglisi F, de Azambuja E, El Saghir NS, Awada A. Angiogenesis and cancer: A cross-talk between basic science and clinical trials (the "do ut des" paradigm). Crit Rev Oncol Hematol 2006;59:40–50.
37. Mancuso A, Sternberg CN. Colorectal cancer and antiangiogenic therapy: what can be expected in clinical practice? Crit Rev Oncol Hematol 2005;55:67–81.
38. Muller AJ, Scherle PA. Targeting the mechanisms of tumoral immune tolerance with small-molecule inhibitors. Nat Rev Cancer 2006;6:613–25.
39. Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. Nucleic Acids Res 2002;30:412–5.
40. Zheng CJ, Han LY, Yap CW, et al. Therapeutic targets: progress of their exploration and investigation of their characteristics. Pharmacol Rev 2006;58:259–79.
41. Irish JM, Kotecha N, Nolan GP. Mapping normal and cancer cell signalling networks: towards single-cell proteomics. Nat Rev Cancer 2005;6:146–55.
42. Furey TS, Cristianini N, Duffy N, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16:906–14.
43. Guo L, Ma Y, Ward R, et al. Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma. Clin Cancer Res 2006;12:3344–54.
44. Edgerton E, Fisher H, Tang L, Frey LJ, Chen Z. Data mining for gene networks relevant to poor prognosis in lung cancer via backward-chaining rule induction. Cancer Informatics 2007;2:93–114.
45. Tang EK, Suganthan PN, Yao X. Gene selection algorithms for microarray data based on least squares support vector machine. BMC Bioinformatics 2006;7:95.

Derivation of Stable Microarray Cancer-Differentiating Signatures Using Consensus Scoring of Multiple Random Sampling and Gene-Ranking Consistency Evaluation

Zhi Qun Tang, Lian Yi Han, Hong Huang Lin, et al.

Cancer Res 2007;67:9996-10003.