JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 10, 2011

Pp. 1329–1338
DOI: 10.1089/cmb.2010.0085

Biomarker Discovery Using Statistically Significant Gene Sets

HOON KIM, JOHN WATKINSON, and DIMITRIS ANASTASSIOU

ABSTRACT

Analysis of large gene expression data sets in the presence and absence of a phenotype can lead to the selection of a group of genes serving as biomarkers jointly predicting the phenotype. Among gene selection methods, filter methods derived from ranked individual genes have been widely used in existing products for diagnosis and prognosis. Univariate filter approaches selecting genes individually, although computationally efficient, often ignore gene interactions inherent in the biological data. On the other hand, multivariate approaches selecting gene subsets are known to have a higher risk of selecting spurious gene subsets due to the overfitting of the vast number of gene subsets evaluated. Here we propose a framework of statistical significance tests for multivariate feature selection that can reduce the risk of selecting spurious gene subsets. Using three existing data sets, we show that our proposed approach is an essential step to identify such a gene set that is generated by a significant interaction of its members, even improving classification performance when compared to established approaches. This technique can be applied for the discovery of robust biomarkers for medical diagnosis.

Key words: cancer classification, gene expression, gene interaction, gene selection, microarray.

1. INTRODUCTION

THERE IS AN INCREASING AMOUNT OF HIGH-THROUGHPUT BIOLOGICAL DATA AVAILABLE, including gene and protein expression, genetic variation, phosphoproteomics, and various forms of imaging signals. Often these data sets are associated with two or more phenotypes, allowing for the discovery of biomarkers that have significant medical utility for diagnosis, prognosis and treatment selection (Baker, 2005; Frank and Hargreaves, 2003). In particular, panels of multiple bio-markers from gene expression data have been proposed and are used in existing products (Paik et al., 2004; van 't Veer et al., 2002). Because of the large number of features and the high degree of noise in high-throughput biological data sets, it is necessary to select a subset of features that are relevant to the phenotype in question (Guyon, 2003).

Feature selection methods for classification can be widely organized into three categories, depending on how they interact with the construction of the classification model (Saeys et al., 2007). Filter

Center for Computational Biology and Bioinformatics, Department of Electrical Engineering, Columbia University, New York, New York.

methods evaluate the association of features with the phenotype of interest independently of the classification model, looking only at the intrinsic characteristics of data. Most filter approaches employ a criterion to evaluate the association of each gene individually. We refer to this approach as univariate feature selection (Baldi and Long, 2001; Ben-Dor et al., 2000; Bhattacharyya et al., 2003; Breitling et al., 2004; Fox and Dimmic, 2006; Golub et al., 1999; Jaeger et al., 2003; Newton et al., 2001; Thomas et al., 2001; Troyanskaya et al., 2002; Zhang et al., 2006). In all cases, the univariate approach is based on including the highest-ranked individual features depending on a chosen association measure, for example, signal-to-noise ratio (Golub et al., 1999). Unlike filter methods, wrapper and embedded methods perform feature selection by using a specific classification model. Wrapper methods involve combinatorial searches through the space of feature subsets, guided by the prediction ability of a specific classification model (Inza et al., 2000; Jirapech-Umpai and Aitken, 2005; Li et al., 2001; Ooi and Tan, 2003; Xiong et al., 2001). On the other hand, embedded methods perform feature selection in the process of training a specific classification model (Diaz-Uriarte and Alvarez de Andres, 2006; Guyon et al., 2002; Jiang et al., 2004; Ma and Huang, 2005). A representative embedded method is Recursive Feature Elimination using Support Vector Machines (SVM-RFE) (Guyon et al., 2002), which uses successive elimination of individual features ranked lowest according to a criterion, aimed at keeping the discrimination ability as high as possible.

Among feature selection methods for classification, univariate filter methods have been dominant in this field because of its simplicity and efficiency. However, it does not take into account gene-gene interactions, possibly leading to less accurate classifiers. Thus, multivariate filter techniques that try to capture the correlations between genes have been proposed (Bo and Jonassen, 2002; Gevaert et al., 2006; Mamitsuka, 2006; Peng et al., 2005; Wang et al., 2005; Xing et al., 2001; Yeoh et al., 2002). Rather than selecting only from a list of highly ranked individual features, multivariate filter techniques, such as feature selection by mutual information can exploit synergistic feature interactions. For example, analysis of a rich training dataset of gene expression data can reveal high correlations of a gene pair with a phenotype, while the individual associations of one or both of the two genes in the pair may not be strong (Anastassiou, 2007), so these genes would not individually appear as highly ranked. Including such gene pairs can improve performance of biomarker classifier. The idea of exploiting gene interactions for feature selection first appeared in Bo and Jonassen (2002), where a subset of gene pairs was selected based on a criterion that involved linear discriminant analysis and a two-sample t-test. While confirming the claim that class prediction can be improved using gene pairs, their approach has a weakness that some genes pairs can be selected just by chance, as pointed out by the authors, because of the huge number of pairs evaluated.

A criterion is required to filter out such spurious gene pairs. In most cases, the statistical significance of some measure of association between each gene pair and the phenotype of interest is assessed, resulting in only those gene pairs that are significant above a predefined significance level being selected. We refer to this test as the overall significance test. However, while this test is proper if it is used for assessing genes individually, it is not sufficient to test genes in pairs, because this association can be generated by an insignificant paring of two genes, in which case just one of the two genes might play a significant role in the association.

Addressing this question in this work, we propose a multivariate feature selection method that identifies small gene sets, such as pairs, each predicting the phenotype in a statistically significant manner. Using several real examples in existing data sets, we demonstrate that our proposed approach is a necessary step to filter out the small gene sets that can occur just by chance. We also show that it can identify many potentially important gene interactions that are associated with the phenotype of interest, revealing that the selected gene sets share a common biological relationship. Regarding classification, any standard classification model can be implemented using the selected small gene sets. In this work, we are interested in a model for which an interpretation of classification decision is simple. Our proposed classification model consists of building such a classifier based on multiple ''voters,'' each being a small statistically significant set of selected genes. We demonstrate that this technique can improve classification performance when compared with three published approaches: correlation-based approach (van 't Veer et al., 2002) from univariate filter methods, gene-pair selection ignoring the statistical significance from multivariate filter methods (Bo and Jonassen, 2002), and SVM-RFE (Guyon et al., 2002) from embedded methods. We hope that it can lead to improved products using gene expression data for medical diagnosis and prognosis.

2. GENE EXPRESSION DATA SETS

For testing our method, we used three publicly available microarray data sets: one containing expression levels of prostate cancer samples, one of colon cancer samples, and one of hepatocellular carcinoma samples. For the prostate cancer set (Singh et al., 2002), raw probe data for a set of Affymetrix Human Genome U95Av2 microarray assays were downloaded from the Broad Institute's website, then normalized using the Robust Multi-array Average (RMA) method (Irizarry et al., 2003), which was implemented in an R package from Bioconductor (Gentleman et al., 2004). The set consists of gene expression profiles of 12,625 genes from 52 tumor samples and 50 normal samples. The colon cancer data set (Alon et al., 1999) consists of gene expression profiles of 2,000 genes from 40 tumor samples and 22 normal samples, originating from Affymetrix Hum6000 arrays. The hepatocellular carcinoma (HCC) data set consists of gene expression profiles of 10,404 genes that are measured using cDNA microarray from pair-matched tumor and adjacent liver tissues from 56 HCC patients. The data set is available from the Gene Expression Omnibus's website (GEO accession number GSE14811).

3. GENE SELECTION AND CLASSIFICATION MODEL

To build an algorithm based on our approach, we need to use a measure estimating the association between a subset of genes and the phenotype. In this work, this is simply measured by the error rate of a linear classifier based on the genes, resulting from a Fisher linear discriminant finding a linear decision boundary orthogonal to the projection line on which the ratio of the between-class variance to the within-class variance is maximized. It should be noted that any measure of association such as t-test performed on the points projected on the linear discriminant axis (Bo and Jonassen, 2002) can be plugged into our approach for gene selection. We also need to introduce a test of the statistical significance of the value of the association measure on the training set found for a particular subset of genes under consideration for being selected. As described in a subsequent section, the significance is ascertained by random permutation testing, in which the association measure found on the training set is compared with the measures calculated from randomly label-permuted versions of the training set.

We first describe the significance testing for gene pairs: Each pair is deemed a ''legitimate voter'' if it passes both of the following two significance tests:

(1) The ''overall significance test,'' which examines whether or not the observed correlation of the pair with the class label in the training set can be due to pure chance.
(2) The ''incremental significance test,'' which compares the correlation of the pair with the correlation of the most discriminatory gene (the ''main'' gene) between the two member genes, examining whether or not the observed improvement following the addition of the ''partner'' gene can be due to pure chance.

3.1. Overall significance test

For the overall significance test, we generated 100 data sets by randomly permuting the class labels of the samples in the original data set. For each permutation we find amongst all possible gene pairs the single pair that achieves the minimum error rate and record that error rate. In our work, the threshold of acceptable significance is defined as the overall minimum error rate over all permutations; in other words, only those pairs whose error rates were never achieved in all 100 permutations were accepted as being overall significant. For the incremental significance test described in a subsequent section, we also generated 100 permutations as above except that the expression values of the main gene remain associated with the true class label for each sample, thus assuring that the correlation of the main gene with the class label is not modified. For each permutation we find the partner gene to the main gene that minimizes the overall error rate and record that error rate. As before, only those pairs whose error rates were never achieved in all permutations were accepted as being incrementally significant.

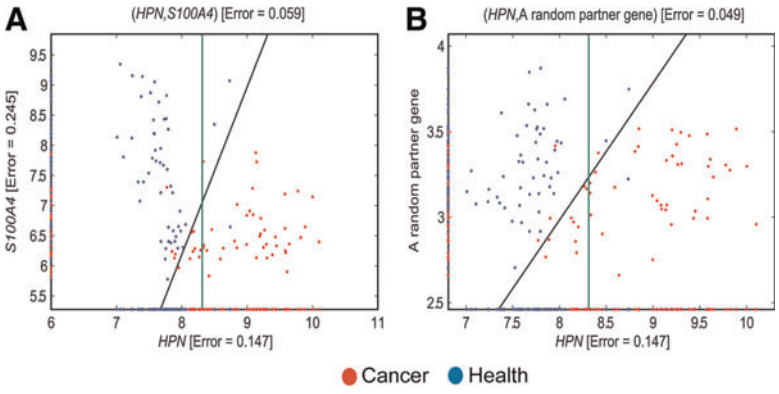

FIG. 1. Scatter plots including the highly discriminating gene *HPN* (hepsin) in the prostate cancer data set. Cancerous samples correspond to red dots, and healthy samples correspond to blue dots. The horizontal axis shows the expression values of *HPN*. The green line indicates the best *HPN*-based classifier, while the black line indicates the best gene-pair-based classifier. Each axis also shows the corresponding error rates. **(A)** Shown in the vertical axis is *S100A4*, the best partner gene to *HPN*, decreasing the error rate to 0.059 for the pair (*HPN*, *S100A4*) compared to 0.147 for *HPN* alone. **(B)** Shown in the vertical axis is the best fictitious partner gene to *HPN* resulting from one of the random permutations described in the text, decreasing the error rate to 0.049 for the pair.

3.2. *Incremental significance test*

The incremental significance test is needed because we must ascertain that each of the two genes, not just one of the two, plays a significant role in the observed high correlation of the pair, so that the addition of the partner gene enhances performance in a statistically significant manner over that of the main gene. For example, it is known that hepsin (HPN) has exceptionally good classification performance in prostate cancer (Magee et al., 2001), which is confirmed by analyzing the data set we used, in which HPN was found as the most discriminatory single gene. Using exhaustive search, we found that S100A4 is the best partner gene to HPN, decreasing the error rate to 0.059 for the pair (HPN, S100A4) compared to 0.147 for HPN alone (Fig. 1A). While the pair passes the overall significance test, it fails to pass the incremental significance test, because the improvement in error rate of the pair does not exceed that which can be attained by pairing the best permuted genes with HPN. As an illustrating example, when HPN was augmented by the best fictitious partner gene resulting from one of the random permutations described above, the error rate was even lower (0.049; Fig. 1B). Interestingly, S100A4 has been implicated in prostate cancer (Saleem et al., 2006), which explains why its individual error rate is found relatively low (0.245), and these results indicate that HPN and S100A4 may be used as individual voters, but we should not use their combination as a single voter.

In contrast, Figure 2 shows two pairs of genes that passed both significance tests. The first pair (Fig. 2A) from the prostate cancer data set contains genes RBP1 and EEF1B2, a pair that was also found using a different methodology and a potential biological explanation was presented (Watkinson et al., 2008). The second pair (Fig. 2B) from the colon cancer data set provides an example of two genes that are not good classifiers individually, but in combination are good and significant predictors of the phenotype. In datasets with very large number of samples, the above testing procedure generalizes to larger gene sets, such as gene

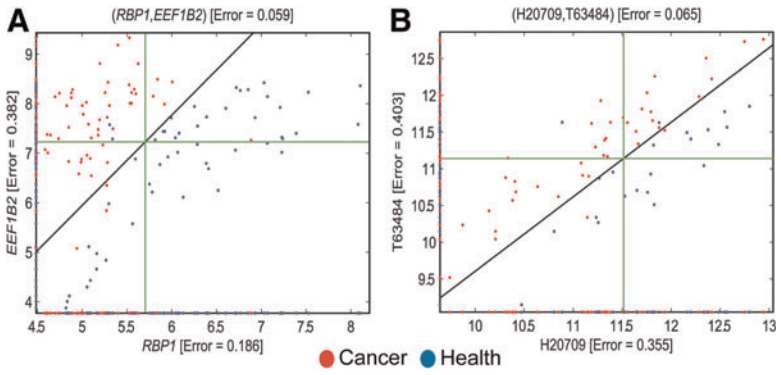

FIG. 2. Examples of scatter plots of significant gene pairs. **(A)** Scatter plot of two genes (*RBP1, EEF1B2*) in the prostate cancer data set that, in combination, significantly improves discrimination ability when compared with each gene alone. **(B)** Scatter plot of two transcripts in the colon cancer data that, in combination, significantly improves discrimination ability compared to each gene alone.

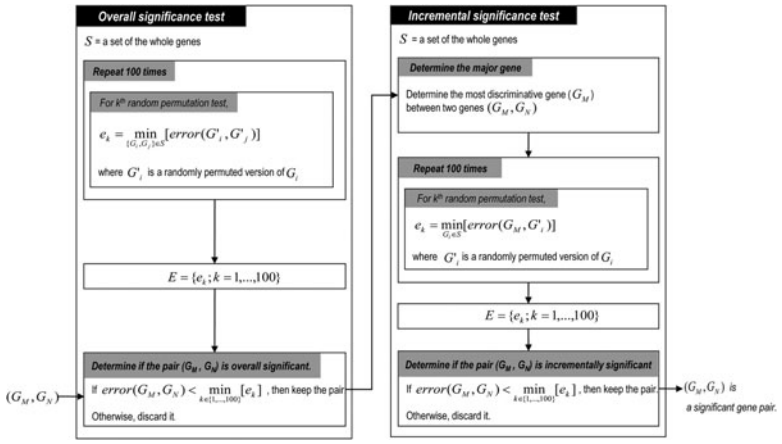

FIG. 3. The significance test protocol. The input is a pair of genes, and the output is an indicator of a significant gene pair. Here, we assume that *GM* is the main gene, the most discriminate gene of the pair (*GM*, *GN*). Only those gene pairs passing both significance tests were considered as significant.

triplets. For example, the correlation of a gene triplet is compared with the correlation of the minimum-error-rate gene pair subset, to ascertain whether the addition of the third gene improves performance in a statistically significant manner.

3.3. Gene set ranking

In our gene selection method, we first filter all gene pairs so that those that do not pass the two significance tests described above are excluded (Fig. 3). We then rank-order all remaining significant gene pairs. We stop when we have used the allocated number of genes. It is in principle (using heuristic search) possible to also identify significant triplets of genes to include in the list. In our examples, however, we found that all triplets were far from being significant following long heuristic search. Each of the selected entries of the list defines an individual linear classifier, serving as a voter. The final aggregate classifier combining together the individual entries of the list is simple majority voting of the individual classifiers. If the number of voters is even and there are an equal number of voters for each class label, then classification results from the selection of the highest-ranked voter.

4. RESULTS

4.1. Classification performance

We compared the classification performance of our proposed method with those from the correlation coefficient based classifier used in van 't Veer et al. (2002), linear SVM classifier combined with linear SVM-RFE used in Guyon et al. (002), and majority voting of individual classifiers defined by error-ranked gene pairs (majority voting with error-ranked gene pairs). In linear SVM classifier combined with linear SVM-RFE, to optimize performance, at each elimination step we removed a single gene, rather than a group of genes. We did two fivefold cross-validation tests to evaluate the performances of different methods. For each method, the entire process—including gene selection and building a classifier—was performed with each training set, and we recorded the error of the trained classifier in the corresponding testing set, calculated as the percentage of wrong predictions (number of wrong predictions divided by the number of the tested samples). Upon completion of all cross-validation experiments, we evaluated the average error versus the number of genes used. Results from our proposed method, correlation-coefficient based classifier, majority voting with error-ranked gene pairs, linear SVM-RFE are shown for the prostate cancer data set, the colon cancer data set, and the hepatocellular carcinoma data set (Fig. 4). For all of three data sets, we found many significant gene pairs in each training set, and we observed that our proposed method outperforms both linear SVM-RFE classifier and correlation-based classifier in all cases. We also observed that our approach using significant gene sets performs better than majority voting with error-ranked gene pairs for both colon cancer and hepatocellular carcinoma data sets, establishing the importance of using the significance tests (less than 32% of the gene pairs selected from an error-ranked list were significant) in both colon cancer and hepatocellular carcinoma data sets. However, this behavior was not

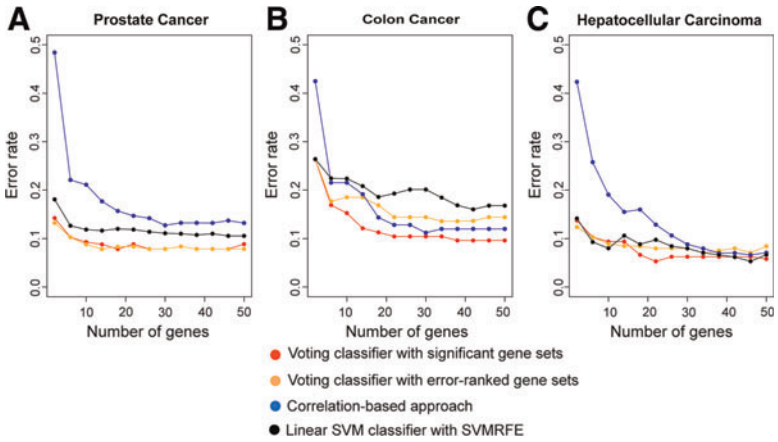

FIG. 4. Classification results for three microarray data sets. The classification error rates of the majority voting method using significant gene sets is compared with those of different methods on the prostate cancer data set **(A),** the colon cancer data set **(B),** and the hepatocellular carcinoma data set **(C),** respectively. In all cases, the vertical axis shows the classification error rate, and the horizontal axis shows the gene subset size.

observed in the prostate cancer data set as more than 82 of the top 100 gene pairs passed the significance test anyway.

4.2. Selected gene sets

As illustrated in Figure 2, selected gene pairs are often significantly associated with phenotypes of interests only when combined together, while the individual members are only weakly associated with the phenotype, suggesting the presence of synergy (Anastassiou, 2007). Since these pairs are incrementally significant, their interaction with respect to the phenotype may be biologically relevant to the study of the disease.

In the prostate cancer data set, the genes that appear in many of the significant gene pairs (Table 1) are RBP1 (cellular retinol-binding protein-1) and DF (human adipsin complement factor D), consistent with the results in Watkinson et al. (2008). In all gene pairs that include RBP1 and DF, the molecular logic is that prostate cancer tends to occur when RBP1 (or DF) is down-regulated, and its partner gene (often a ribosomal protein) is up-regulated. Many of the genes in the significant pairs are ranked low in the list of

TABLE 1. RANKING OF SIGNIFICANT GENE PAIRS BY LINEAR ERROR RATE IN THE PROSTATE CANCER DATA SET

Ranking	*Accession 1*	*Symbol 1*	*Accession 2*	*Symbol 2*	*Linear error rate*
1	AC002400	*NDUFAB1*	M11433	*RBP1*	0.049
2	X60489	*EEF1B2*	M11433	*RBP1*	0.0588
3	M98539	*PTGDS*	U28686	*RBM3*	0.0588
4	AA149307	*TCEAL4*	J03592	*SLC25A6*	0.0588
5	U68063	*SFRS10*	M84526	*DF*	0.0588
6	Z93930	*XBP1*	M84526	*DF*	0.0588
7	M29039	*JUNB*	M84526	*DF*	0.0588
8	D86640	*STAC*	AJ130733	*AMACR*	0.0588
9	J03592	*SLC25A6*	M11433	*RBP1*	0.0588
10	Z25749	*RPS7*	M11433	*RBP1*	0.0686
11	AA426364	*ATP5I*	M11433	*RBP1*	0.0686
12	U14972	*RPS10*	M84526	*DF*	0.0686
13	X16416	*ABL1*	M11433	*RBP1*	0.0686
14	X56681	*JUND*	M84526	*DF*	0.0686
15	M32304	*TIMP2*	M84526	*DF*	0.0686
16	AA135683	*BASP1*	M11433	*RBP1*	0.0686
17	AF068179	*CAMLG*	M84526	*DF*	0.0686
18	AI961040	*TUBGCP2*	M84526	*DF*	0.0686
19	AF065388	*TSPAN1*	M84526	*DF*	0.0686
20	X56932	*RPL13A*	M84526	*DF*	0.0686

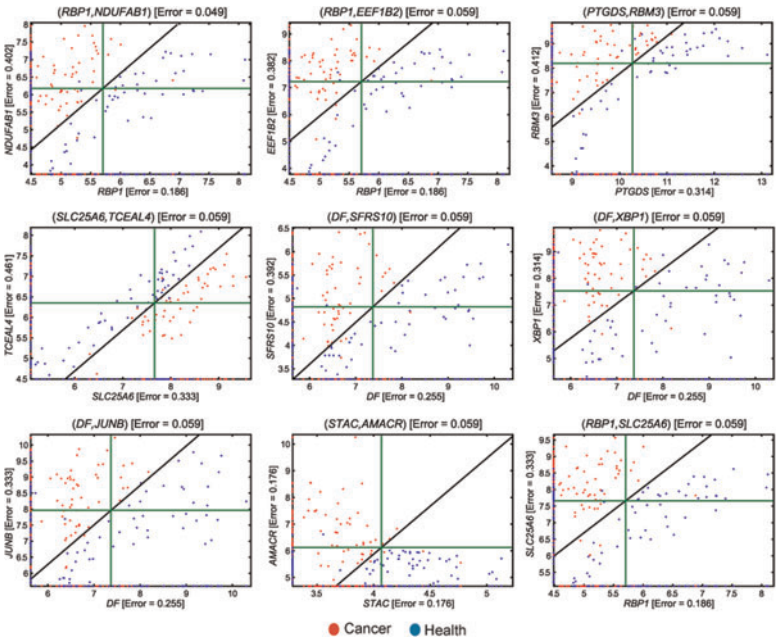

FIG. 5. The scatter plots of nine significant gene pairs in the prostate cancer data set.

single genes and would be missed by traditional approaches (Fig. 5). For example, the cancer-related gene, RBM3 (Sureban et al., 2008), is a low-ranked single gene, but is a member of a high-ranked significant gene pair.

In the hepatocellular carcinoma data set, RPL8 (ribosomal protein L8) appears in many of the significant gene pairs (Table 2). It is interesting that all the gene pairs that include RPL8 also follow the same molecular logic; in hepatocellular carcinoma tissues, RPL8 is expressed at high levels, while its partner gene is expressed at low levels (Fig. 6). It is known that RPL8 is over-expressed in several tumors, having potential as a vaccine target for those tumors (Swoboda et al., 2007).

Table 2. Ranking of Gene Pairs by Linear Error Rate in the Hepatocellular Carcinoma Data Set

Ranking	*Accession 1*	*Symbol 1*	*Accession 2*	*Symbol 2*	*Linear error rate*
1	AF290475	*RSBN1L*	NM_033301	*RPL8*	0.044643
2	NM_002346	*LY6E*	NM_033301	*RPL8*	0.053571
3	NM_014888	*FAM3C*	NM_033301	*RPL8*	0.0625
4	BI488702	*MT1M*	NM_003122	*SPINK1*	0.0625
5	NM_002346	*LY6E*	NM_001235	*SERPINH1*	0.071429
6	NM_020992	*PDLIM1*	NM_033301	*RPL8*	0.071429
7	NM_002268	*KPNA4*	NM_002568	*PABPC1*	0.071429
8	NM_005410	*SEPP1*	NM_033301	*RPL8*	0.071429
9	NM_014888	*FAM3C*	NM_003122	*SPINK1*	0.071429
10	NM_001920	*DCN*	NM_000050	*ASS1*	0.071429
11	NM_002346	*LY6E*	BI768064	*RAB34*	0.071429
12	NM_002346	*LY6E*	NM_002863	*PYGL*	0.071429
13	BF131637	*MT2A*	NM_003122	*SPINK1*	0.071429
14	NM_014427	*CPNE7*	NM_033301	*RPL8*	0.071429
15	BG749845	*MT1E*	NM_033301	*RPL8*	0.080357
16	NM_005836	*HRSP12*	NM_033301	*RPL8*	0.080357
17	NM_006007	*ZFAND5*	NM_033301	*RPL8*	0.080357
18	NM_002268	*KPNA4*	NM_033301	*RPL8*	0.080357
19	BG749845	*MT1E*	NM_002568	*PABPC1*	0.080357
20	NM_005836	*HRSP12*	NM_015658	*NOC2L*	0.080357

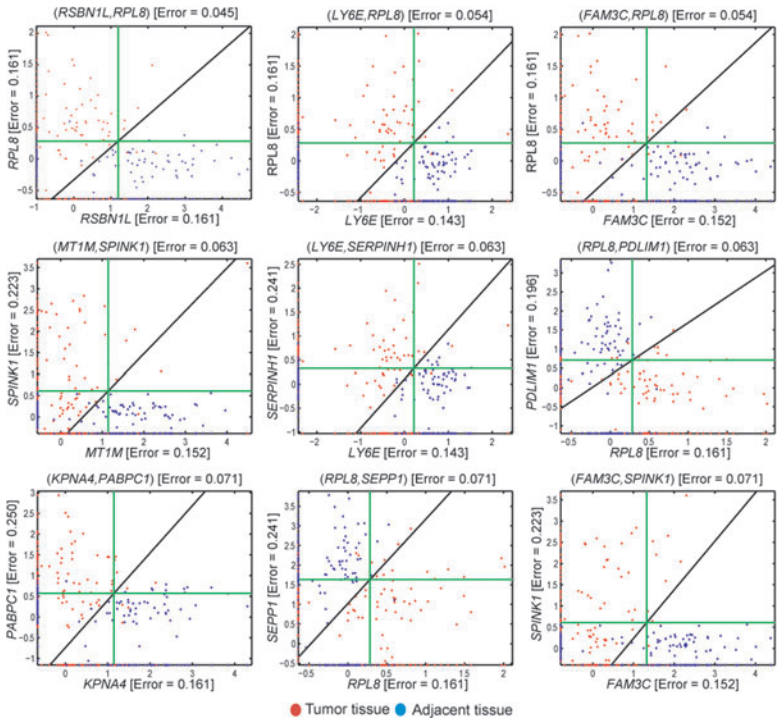

FIG. 6. The scatter plots of nine significant gene pairs in the hepatocellular data set.

5. CONCLUSION

The goal of using statistical significance tests for feature selection is that it can reduce a risk of selecting irrelevant genes that might originate from random fluctuations in biological systems. Our approach is the logical extension of the statistical tests in a multivariate way, which is to find genes that play a significantly cooperative role in predicting the phenotype. By addressing the statistical significance of the error rate of gene sets, we can reduce the risk of selecting irrelevant features. Using several real examples, we demonstrated that it is an essential step for multivariate feature selection. Our approach can be further extended to identifying other types of genes by using other scoring measures such as mutual information. Although microarray data are known to be noisy and platform dependent, a potential single product measuring expression levels under specified and well-controlled conditions will guarantee the reliability of the biomarker classifier, as in existing commercial medical diagnostic products. Our algorithm can also naturally be applied to other types of biomarkers—such as imaging-based signals, SNPs, Copy Number variations, and phosphoproteomics—including mixtures of biomarkers from diverse types. We hope that our approach will provide a valuable computational tool helpful for the design of accurate and robust biomarker products for medical diagnosis and prognosis.

ACKNOWLEDGMENTS

We thank Vinay Varadan for helpful discussions.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Alon, U., Barkai, N., Notterman, D.A., et al. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96, 6745–6750.

Anastassiou, D. 2007. Computational analysis of the synergy among multiple interacting genes. *Mol. Syst. Biol.* 3, 83.
Baker, M. 2005. In biomarkers we trust? *Nat. Biotechnol.* 23, 297–304.
Baldi, P., and Long, A.D. 2001. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17, 509–519.
Ben-Dor, A., Bruhn, L., Friedman, N., et al. 2000. Tissue classification with gene expression profiles. *J. Comput. Biol.* 7, 559–583.
Bhattacharyya, C., Grate, L., Rizki, A., et al. 2003. Simultaneous classification and relevant feature identification in high-dimensional spaces: application to molecular profiling data. *Signal Process.* 83, 729–743.
Bo, T., and Jonassen, I. 2002. New feature subset selection procedures for classification of expression profiles. *Genome Biol.* 3, research0017.
Breitling, R., Armengaud, P., Amtmann, A., et al. 2004. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* 573, 83–92.
Diaz-Uriarte, R., and Alvarez de Andres, S. 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7, 3.
Fox, R.J., and Dimmic, M.W. 2006. A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics* 7, 126.
Frank, R., and Hargreaves, R. 2003. Clinical biomarkers in drug discovery and development. *Nat. Rev. Drug Discov.* 2, 566–580.
Gentleman, R.C., Carey, V.J., Bates, D.M., et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5, R80.
Gevaert, O., De Smet, F., Timmerman, D., et al. 2006. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* 22, e184-e190.
Golub, T.R., Slonim, D.K., Tamayo, P., et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.
Guyon, I. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.
Guyon, I., Weston, J., Barnhill, S., et al. 2002. Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46, 389–422.
Inza, I., Larrañaga, P., Etxeberria, R., et al. 2000. Feature subset selection by Bayesian network-based optimization. *Artif. Intell.* 123, 157–184.
Irizarry, R.A., Hobbs, B., Collin, F., et al. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264.
Jaeger, J., Sengupta, R., Ruzzo, W.L., et al. 2003. Improved gene selection for classification of microarrays. *Pac. Symp. Biocomput*. 53–64.
Jiang, H., Deng, Y., Chen, H.-S., et al. 2004. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* 5, 81.
Jirapech-Umpai, T., and Aitken, S. 2005. Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. *BMC Bioinformatics* 6, 148.
Li, L., Weinberg, C.R., Darden, T.A., et al. 2001. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics* 17, 1131–1142.
Ma, S., and Huang, J. 2005. Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics* 21, 4356–4362.
Magee, J.A., Araki, T., Patil, S., et al. 2001. Expression profiling reveals hepsin overexpression in prostate cancer. *Cancer Res.* 61, 5692–5696.
Mamitsuka, H. 2006. Selecting features in microarray classification using ROC curves. *Pattern Recognit.* 39, 2393–2404.
Newton, M.A., Kendziorski, C.M., Richmond, C.S., et al. 2001. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.* 8, 37–52.
Ooi, C.H., and Tan, P. 2003. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics* 19, 37–44.
Paik, S., Shak, S., Tang, G., et al. 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* 351, 2817–2826.
Peng, H.C., Ding, C., and Long, F., et al. 2005. Minimum redundancy—maximum relevance feature selection. *IEEE Intell. Syst.* 20, 70–71.
Saeys, Y., Inza, I., and Larrañaga, P. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 2507–2517.
Saleem, M., Kweon, M.-H., Johnson, J.J., et al. 2006. S100A4 accelerates tumorigenesis and invasion of human prostate cancer through the transcriptional regulation of matrix metalloproteinase 9. *Proc. Natl. Acad. Sci. USA* 103, 14825–14830.
Singh, D., Febbo, P.G., Ross, K., et al. 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1, 203–209.

Sureban, S.M., Ramalingam, S., Natarajan, G., et al. 2008. Translation regulatory factor RBM3 is a proto-oncogene that prevents mitotic catastrophe. *Oncogene* 27, 4544–4556.

Swoboda, R.K., Somasundaram, R., Caputo, L., et al. 2007. Shared MHC class II--dependent melanoma ribosomal protein L8 identified by phage display. *Cancer Res.* 67, 3555–3559.

Thomas, J.G., Olson, J.M., Tapscott, S.J., et al. 2001. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.* 11, 1227–1236.

Troyanskaya, O.G., Garber, M.E., Brown, P.O., et al. 2002. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* 18, 1454–1461.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536.

Wang, Y., Tetkoa, I.V., Hall, M.A., et al. 2005. Gene selection from microarray data for cancer classification—a machine learning approach. *Comput. Biol. Chem.* 29, 37–46.

Watkinson, J., Wang, X., Zheng, T., et al. 2008. Identification of gene interactions associated with disease from gene expression data using synergy networks. *BMC Syst. Biol.* 2, 10.

Xing, E.P., Jordan, M.I., Karp, R.M., et al. 2001. Feature selection for high-dimensional genomic microarray data. *Proc. 18th Int. Conf. Mach. Learn.* Pages 601–608.

Xiong, M., Fang, X., and Zhao, J. 2001. Biomarker identification by feature wrappers. *Genome Res.* 11, 1878–1887.

Yeoh, E.-J., Ross, M.E., Shurtleff, S.A., et al. 2002. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 1, 133–143.

Zhang, C., Lu, X., and Zhang, X. 2006. Significance of gene ranking for classification of microarray samples. *IEEE/ACM Trans Comput. Biol. Bioinform.* 3, 312–320.

Address correspondence to:
Dr. Dimitris Anastassiou
Center for Computational Biology and Bioinformatics
Department of Electrical Engineering
Columbia University
New York, NY 10027

E-mail: anastas@ee.columbia.edu