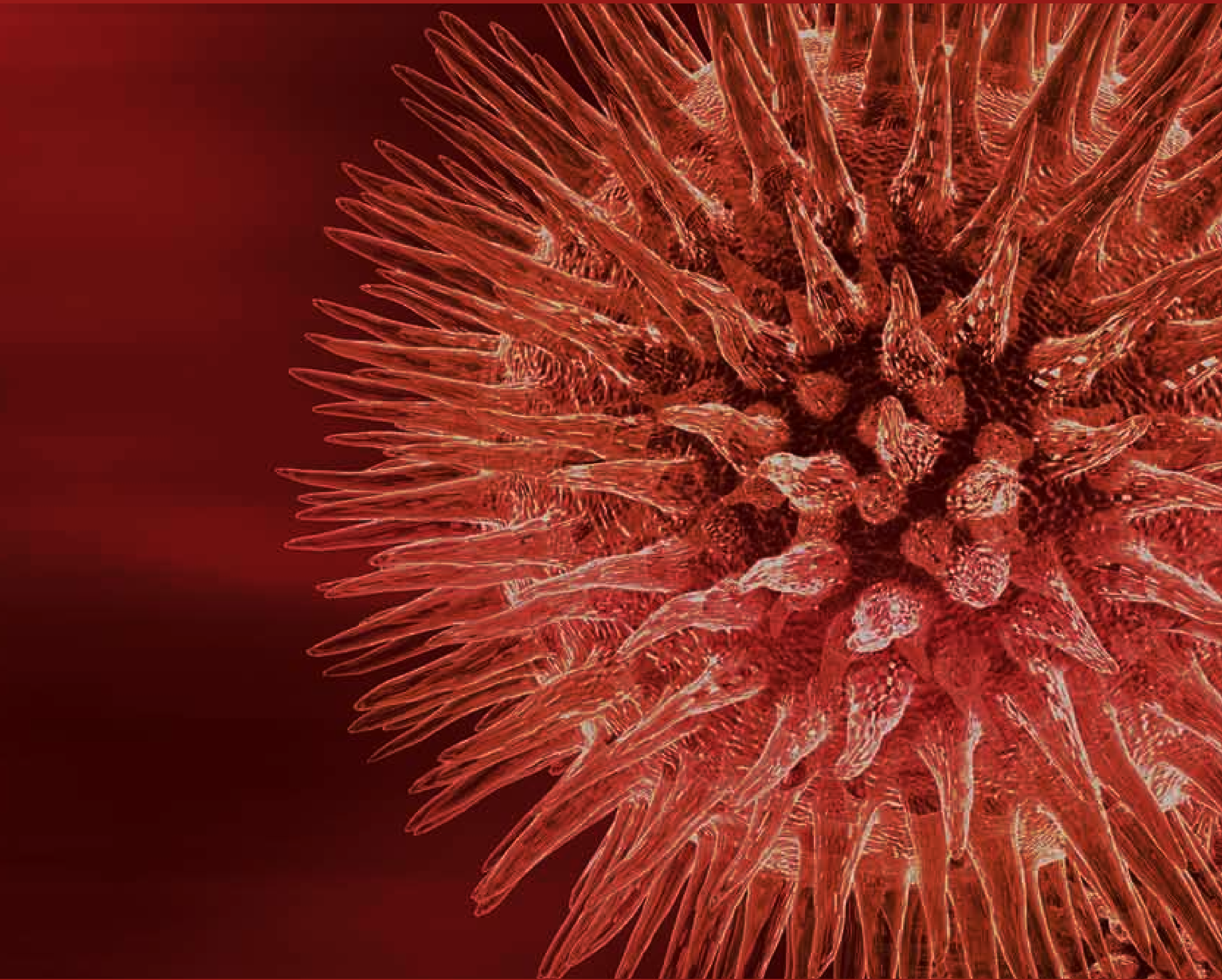
Journal of Biomedicine and Biotechnology

Computational Data Mining in Cancer Bioinformatics and Cancer Epidemiology

Guest Editors: Zhenqiu Liu, Dechang Chen, Xuewen Chen, and Haomiao Jia



Computational Data Mining in Cancer Bioinformatics and Cancer Epidemiology

Journal of Biomedicine and Biotechnology

Computational Data Mining in Cancer Bioinformatics and Cancer Epidemiology

Guest Editors: Zhenqiu Liu, Dechang Chen,
Xuewen Chen, and Haomiao Jia

Editorial Board

The editorial board of the journal is organized into sections that correspond to the subject areas covered by the journal.

Agricultural Biotechnology

Animal Biotechnology

Biochemistry

Bioinformatics

Biophysics

Cell Biology

Genetics

Genomics

Immunology

Microbial Biotechnology

Microbiology

Molecular Biology

Oncology

Pharmacology

Plant Biotechnology

Toxicology

Virology

Contents

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 582697, 2 pages
doi:10.1155/2009/582697

Editorial

Computational Data Mining in Cancer Bioinformatics and Cancer Epidemiology

Zhenqiu Liu,[1] Dechang Chen,[2] Xuewen Chen,[3] and Haomiao Jia[4]

[1] *Division of Biostatistics and Bioinformatics, The University of Maryland Greenebaum Cancer Center, USA*
[2] *Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, USA*
[3] *Department of Electrical Engineering and Computer Science, The University of Kansas, USA*
[4] *Mailman School of Public Health, Columbia University, USA*

Correspondence should be addressed to Zhenqiu Liu, zliu@umm.edu

Received 18 August 2009; Accepted 18 August 2009

High throughput technologies such as microarray have produced huge amount of genomic and proteomic data in public domain. Many survey and clinical outcome data such as SEER data are also available. A long list of links to large health-related data sets can be found at the website http://www.ehdp.com/. All of these databases have different temporal and spatial assumptions for example, different frequencies of collection, different spatial resolution (by state, by county, by zip-code, and by square kilometer), and so forth. How to mine these data together and extract useful information is really a challenging task. This special issue brings together researchers from different disciplines and encourages collaborative research on cancer related computational data mining. The objectives of this special issue are intended to address two challenging issues. One is how to identify and evaluate biomarkers (features, risk/protector) factors. The other is to develop new or adapt existing algorithms to analyze data from different sources.

This special issue published 11 articles, which may be classified into three groups: (1) those concerned with problems with gene selection and predictions, (2) those developed methods for network construction and system biology with multi source genomic data, and (3) those related to medical informatics and methodology research.

The first group covers methods in gene selection and prediction, a fundamental problem in biomedical research. These studies open new avenues for identifying complex disease genes and biomarkers for disease diagnosis and for assessing drug efficacy and toxicity. For examples, the l_1 penalized methods can be efficiently implemented with different classifiers for gene identification and model prediction. In one article, Huang and Wu propose a novel method for cancer diagnosis using gene expression data by casting the classification problem as finding sparse representations of test samples with respect to training samples. The sparse representation is computed by the l_1-regularized least square method. The proposed method is more efficient than SVMs as it has no need of model selection. Receiver Operating Characteristic (ROC) analysis is a common tool for assessing the performance of biomarkers and prediction models. It gained much popularity in biomedicine. Liu et al., in another article, propose a novel method through regularized F-measure maximization. The proposed method assigns different costs to positive and negative samples and does simultaneous feature selection and prediction with l_1 penalty. This method is useful especially when data set is highly unbalanced or the labels for negative (positive) samples are missing, which is very common in biomedical research.

Also in the first group, an article by Jrad develops a multiclass cancer diagnosis with class-selective rejection scheme for gene selection. It gives a general formulation of the problem and proposes a possible solution based on ν-1-SVM coupled with its regularization path. The proposed classifier minimizes any asymmetric loss function and consists of rejecting some patients from one, some, or all classes in order to ensure a higher reliability while reducing time and expense costs. Another article on human cancer prediction by Martín-Merino et al. incorporates in the ν-SVM algorithm a linear combination of non-Euclidean dissimilarities. The weights of the combination are learnt in a Hyper Reproducing Kernel Hilbert Space (HRKHS) using a Semidefinite Programming algorithm. This approach allows us to incorporate a smoothing term that penalizes the complexity of the family of distances and avoids overfitting. This method is more robust than the traditional support vector machines (SVMs). Another methodology article by Hua et al. proposes a Bayesian cut fitting to describe features in response to the skeletal age. Their method cannot only capture the entire pattern of feature variation but also carry the local properties regarding the skeletal age.

The second group includes genomic networks and system biology with multi source genomic data. In a biological system, genes perform different molecular functions and regulate various biological processes via interactions with other genes thus forming a variety of complex networks. Article by Han et al. proposes an integrative method based on the bootstrapping K-S test to evaluate a large number of microarray datasets generated from 21 different types of cancer in order to identify gene pairs that have different relationships in normal versus cancer tissues. The significant alteration of gene relations can greatly extend our understanding of the molecular mechanisms of human cancer. In another article, Spencer et al. utilize data mining methods based on machine learning to build a predictive model of lung injury by retrospective analysis of treatment planning archives. In addition, biomarkers for this model are extracted from a prospective clinical trial that collects blood serum samples at multiple time points. They utilize a 3-way proteomics methodology to screen for differentially expressed proteins that are related to RP. They present their proteomic methodology to investigate predictive biomarkers of RP that could eliminate informational gaps in the retrospective physical model. Article by Wang et al. constructs a single gene network based on linear programming and an integrated analysis of the significant function cluster using Kappa statistics and fuzzy heuristic clustering. Finally, in their article, Loganantharaj and Chung introduce an integrating protein-to-protein interaction information, pathway information with array expression data set to identify a set of "important" genes and potential signal transduction networks that help to target and reverse the oncogenic phenotype induced by tumor antigen such as integrin a6b4.

The third group comprises two articles, which cover advances in medical informatics and spatial and temporal data analyses. In their article, Chen et al. develop a prognostic system of cancer patients with ensemble clustering and SEER database. This system can be used to predict an outcome or a survival rate of cancer patients with more accuracy. The article by Song et al. proposes a method of classifying temporal gene expression curves in which individual expression trajectory is modeled as longitudinal data with a changeable variance and covariance structure. The method, mainly based on generalized mixed model, is illustrated by a dense temporal gene expression data in bacteria. The power and time points of measurements are also characterized via the longitudinal mixed model. Even if the method is developed for temporal gene expression data, it may be generally applicable to other spatial and temporal data analyses.

Zhenqiu Liu
Dechang Chen
Xuewen Chen
Haomiao Jia

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 403689, 6 pages
doi:10.1155/2009/403689

Research Article

Sparse Representation for Classification of Tumors Using Gene Expression Data

Xiyi Hang[1] and Fang-Xiang Wu[2, 3]

[1] Department of Electrical and Computer Engineering, California State University, Northridge, CA 91330, USA
[2] Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SK, Canada S7N 5A9
[3] Divsion of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK, Canada S7N 5A9

Correspondence should be addressed to Fang-Xiang Wu, faw341@mail.usask.ca

Received 24 November 2008; Accepted 12 January 2009

Recommended by Dechang Chen

Personalized drug design requires the classification of cancer patients as accurate as possible. With advances in genome sequencing and microarray technology, a large amount of gene expression data has been and will continuously be produced from various cancerous patients. Such cancer-alerted gene expression data allows us to classify tumors at the genomewide level. However, cancer-alerted gene expression datasets typically have much more number of genes (features) than that of samples (patients), which imposes a challenge for classification of tumors. In this paper, a new method is proposed for cancer diagnosis using gene expression data by casting the classification problem as finding sparse representations of test samples with respect to training samples. The sparse representation is computed by the l_1-regularized least square method. To investigate its performance, the proposed method is applied to six tumor gene expression datasets and compared with various support vector machine (SVM) methods. The experimental results have shown that the performance of the proposed method is comparable with or better than those of SVMs. In addition, the proposed method is more efficient than SVMs as it has no need of model selection.

1. Introduction

The treatment of cancer greatly depends on the accurate classification of tumors. In spite of its effectiveness in classifying tumors by microscopic tissue examination, traditional histopathological approach fails to classify many cancer cases. The number of unclassified cancer cases can reach up to 40 000 per year just in the United States [1]. DNA microarray technology, on the other hand, has the potential to provide a more accurate and objective cancer diagnosis due to its high throughput capability of measuring expression levels of tens of thousands genes simultaneously. Since Golub et al. [2] successfully classified between acute myeloid leukemia (AML) and acute lymphocytic leukemia (ALL), many other types of cancer have been classified using gene expression data including breast cancer [3], lymphoma [4], lung cancer [5], bladder cancer [6], colon cancer [7], ovarian cancer [8], prostate cancer [9], melanoma [10], and brain tumors [11].

The successful application of microarray technology in cancer diagnosis greatly depends on the careful design of two important components of a gene data classification system: gene selection and sample classification, shown in Figure 1. Gene selection mainly serves two purposes: (i) to reduce dramatically the number of genes used in classification to manage the "curse of dimensionality" and (ii) selected genes might be biologically relevant, allowing further biological exploration which may lead to better understanding of underlying molecular mechanism associated with tumorigenesis and progression. Gene selection can be made by test statistics [12]. An excellent review on gene selection methods can be found in [13].

The second component, sample classification, is a challenging issue for a problem with a small number of learning samples and yet a large number of features (genes). The number of samples available for analysis ranges from tens to hundreds. Many established methods have been proposed to address the challenge. According to Lee et al. [14], they can

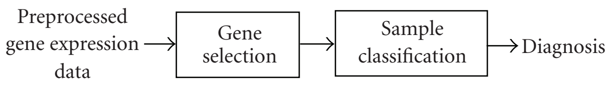


Figure 1: The pipeline of cancer diagnosis using gene expression data.

be classified into four categories: (i) classical methods such as Fisher's linear discriminant analysis, logistic regression, K-nearest neighbor, and generalized partial least square, (ii) classification trees and aggregation methods such as CART, random forest, bagging, and boosting, (iii) machine learning methods such as neural network and support vector machines (SVMs), and (iv) generalized methods such as flexible discriminant analysis, mixture discriminant analysis, and shrunken centroid method.

In this paper, we propose a novel approach for classification, called sparse representation, inspired by the recent progress in l_1-norm minimization-based methods such as basis pursuit denoising [15], compressive sensing for sparse signal reconstruction [16–18], and Lasso algorithm for feature selection [19]. Ideally, a testing sample can be represented just in terms of the training samples of the same category. Hence, when the testing sample is expressed as linear combination of all the training samples, the coefficient vector is sparse, that is, the vector has relatively few nonzero coefficients. Testing samples of same category will have similar sparse representation, while different categories will result in different sparse representations. In order to recover the sparse coefficient vector, l_1-regularized least square [20] is used.

Unlike general supervised learning methods, where a training procedure is used to create a classification model for testing, the sparse representation approach does not contain separate training and testing stages. Instead, classification is achieved directly out of the testing sample's sparse representation in terms of training samples. Another unique feature of the new method is no model selection needed. It is well known that the performance of a classifier, such as SVM, relies upon careful choice of the model parameters via model selection procedure.

2. Materials and Methods

2.1. Sparse Representation. Consider a training dataset $\{(\mathbf{x}_i, l_i);\ i = 1, \dots, n\}$, $\mathbf{x}_i \in R^d$, $l_i \in \{1, 2, \dots, N\}$, where $\mathbf{x}_i$ represents the ith sample, a d-dimensional column vector containing gene expression values with d as the number of genes, and l_i is the label of the ith sample with N as the number of categories. For a testing sample $\mathbf{y} \in R^d$, the problem of sparse representation is to find a column vector $\mathbf{c} = [c_1, c_2, \dots, c_n]^T$ such that

$$\mathbf{y} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n, \tag{1}$$

and $\|c\|_0$ is minimized, where $\|c\|_0$ is l_0-norm, and it is equivalent to the number of nonzero components in the vector $\mathbf{c}$.

Defining a matrix by putting $\mathbf{x}_i$ as the ith column $A = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$, the problem of sparse representation can be converted into

$$\mathbf{c} = \min_{\mathbf{c}' \in R^n} \|\mathbf{c}'\|_0 \quad \text{subject to } \mathbf{y} = A\mathbf{c}. \tag{2}$$

Finding the solution to sparse representation problem is NP-hard due to its nature of combinational optimization. Approximation solution can be obtained by replacing the l_0-norm in (2) by the l_p-norm

$$\mathbf{c} = \min_{\mathbf{c}' \in R^n} \|\mathbf{c}'\|_p \quad \text{subject to } \mathbf{y} = A\mathbf{c}, \tag{3}$$

where the l_p-norm of a vector $\mathbf{v}$ defined as $\|\mathbf{v}\|_p = (\sum_i |v_i|^p)^{1/p}$. A generalized version of (3), which allows for certain degree of noise, is to find a vector $\mathbf{c}$ such that the following objective function is minimized:

$$J(\mathbf{c}, \lambda) = \min_{\mathbf{c}} \{\|A\mathbf{c} - \mathbf{y}\|_2 + \lambda\|\mathbf{c}\|_p\}, \tag{4}$$

where the positive parameter λ is a scalar regularization that balances the tradeoff between reconstruction error and sparsity.

Since l_1-norm minimization can efficiently recover sparse signal [20] and are robust against outliers, this study takes $p = 1$ in (4). Therefore, the problem is reduced to solve (3) an l_1-regularized least square problem:

$$J(\mathbf{c}, \lambda) = \min_{\mathbf{c}} \|A\mathbf{c} - \mathbf{y}\|_2 + \lambda\|\mathbf{c}\|_1. \tag{5}$$

A truncated Newton interior-point method (TNIPM) proposed in [20] can be used to solve the above optimization problem in (5). For the convergence of the algorithm, the regularization parameter must satisfy the following condition:

$$\lambda \leq \|2A^T y\|_\infty. \tag{6}$$

Please refer to [20] for more information about l_1-regularized least square and the specialized interior-point method.

Another approach to determine the sparse solution to (2) is to use the framework of compressive sensing, which requires the system to be underdetermined. Including the construction errors $\mathbf{e}$ in (1) yields

$$\mathbf{y} = A\mathbf{c} + \mathbf{e}. \tag{7}$$

In compressive sensing approach, we need to rewrite (7) as

$$\mathbf{y} = \begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \mathbf{e} \end{bmatrix} = B\mathbf{d}, \tag{8}$$

```
Input: $\{(\mathbf{x}_i, y_i);\ i = 1, \ldots, n\}$, and $\mathbf{y}$
    1. Normalize $\mathbf{x}_i$, $i = 1, 2, \ldots, n$, and $\mathbf{y}$
    2. Create matrix A
    3. Solve the optimization problem defined in (5)
    4. Compute $g_k(\mathbf{y})$, $k = 1, 2, \ldots, N$
Output: $\arg\min_k g_k(\mathbf{y})$
```

Algorithm 1: Classification by sparse representation.

where $\mathbf{B} = [\mathbf{A}\ \ \mathbf{I}] \in R^{d\times(n+d)}$ and $\mathbf{d} = [\mathbf{c}^T\ \ \mathbf{e}^T]^T \in R^{n+d}$. With these notations, the sparse representation can be obtained by the following constrained l_1-norm minimization problem:

$$\min_{\mathbf{d}} \|\mathbf{d}\|_1 \quad \text{subject to } B\mathbf{d} = \mathbf{y}. \tag{9}$$

The above linear programming problem can be solved by a specialized interior-point method called l_1-magic [21]. The approach in (9) is used in [22] for face recognition by sparse representation.

Both approaches do generate nearly the same classification performance in our experiments. Our approach, based on l_1-regularized least square, however, is much faster. First, the optimization problem scale in our approach is much smaller. For example, when the training dataset contains 300 samples and the gene number is 10 000, the matrix in our approach is $A \in R^{10000\times300}$ while $B \in R^{10000\times10300}$. Secondly, TNIPM is $O(n^{1.2})$ while l_1-magic is $O(n^{1.3})$ [20]. In addition, it is noticed that basis pursuit, compressive sensing, and Lasso algorithm can also be converted into l_1-regularized least square problems [20].

Let $\hat{\mathbf{c}}$ denote the sparse representation obtained by l_1-regularized least square. Ideally, the nonzero entries in $\hat{\mathbf{c}}$ are associated with the columns in A corresponding to those training samples of the same category as the testing sample $\mathbf{y}$. However, noises may cause the nonzero entries to be linked with multiple categories [22]. Simple heuristics, such as assigning $\mathbf{y}$ to the category with the largest entry in $\hat{\mathbf{c}}$, are not dependable. Instead, we define N discriminate functions

$$g_k(\mathbf{y}) = \left\|\mathbf{y} - A\hat{\mathbf{c}}_k\right\|_2, \quad k = 1, 2, \ldots, N, \tag{10}$$

where $\hat{\mathbf{c}}_k$ is obtained by keeping only those entries in $\hat{\mathbf{c}}$ associated with category k and assigning zeros to other entries. Thus g_k represents the approximation error when $\mathbf{y}$ is assigned to category k, and we can assign $\mathbf{y}$ to the category with the smallest approximation error. The classification algorithm is summarized (see Algorithm 1).

2.2. Numerical Experiments. Numerical experiments are designed to quantitatively verify the performance of sparse representation method for cancer classification using gene expression data. The performance metric used in this study is accurate, obtained by stratified 10-fold cross-validation. We compare our approach with a few variants of multicategory SVMs. SVMs, as state-of-the-art machine learning algorithms, have been successfully applied in gene profile classification [23, 24]. The comprehensive study in [25] also shows that SVMs outperform K-nearest neighbors and neural network in gene expression cancer diagnosis.

All experiments are done on a PC with duo Intel 2.33 G CPU and 4 G memory under Windows XP (SP2). MATLAB R14 is used to implement sparse representation method. The optimization is done by l1_ls MATLAB package, which is available online (http://www.stanford.edu/~boyd/l1_ls/). The results of SVMs are obtained by gene expression model selector (GEMS), a software with graphic user interface for classification of gene expression data, which is freely available at http://www.gems-system.org/ and used in [25] for the comprehensive study of the performance of multiple classifiers on gene expression cancer diagnosis. Besides standard binary SVM, GEMS has implemented the following multiclass SVMs: one-versus-rest (OVR) [26], one-versus-one (OVO) [26], directed acyclic graph (DAG) [27], all-at-once method by Weston and Watkins (WW) [28], and all-at-once method by Crammer and Singer (CS) [29], which are used in comparison with sparse representation approach. Polynomial and RBF kernels are used for SVMs.

For fair comparison, the partition file of cross-validation generated by GEMS is used in sparse representation approach. As for model selection, 9-fold cross validation is used for SVMs.

The comparison is done with and without gene selection. Two popular gene selection methods are used in this study: Kruskal-Wallis nonparametric one-way ANOVA (KW) [30] and the ratio of between-groups to within-groups sum of squares (BW) [31].

2.3. Datasets. In the experiment, we use six datasets, which are among 11 datasets used in the comprehensive study [25]. For easy comparison, we adopt the name used in [25]. The information about the six datasets is summarized below.

(i) 9_Tumors [32]: the dataset comes from a study of 9 human tumor types: NSCLC, colon, breast, ovary, leukemia, renal, melanoma, prostate, and CNS. There are 60 samples, each of which contains 5726 genes.

(ii) 11_Tumors [23]: the dataset includes 174 samples of gene expression data of 11 various human tumor types: ovary, bladder/ureter, breast, colorectal, gastro-esophagus, kidney, liver, prostate, pancreas, adeno lung, and squamous lung. The number of genes is 12 533.

(iii) 14_Tumors [24]: the dataset contains 308 samples of 14 various human tumor types including leukemia, prostate, lung, colorectal, lymphoma, bladder, melanoma, uterus, breast, renal, pancreas, ovary, mesothelioma, and CNS, and 12 normal tissues including breast, prostate, lung, colon, germinal center, bladder, uterus, peripheral blood, kidney, pancreas, ovary, and brain. Each sample has 15 009 genes.

(iv) Brain_Tumor1 [11]: the dataset comes from a study of 5 human brain tumor types: medulloblastoma, malignant glioma, AT/RT, normal cerebellum, and PNET, including 90 samples. Each sample has 5920 genes.

Table 1: Results without gene selection.

Methods		Prostate_Tumor	9_Tumors	11_Tumors	14_Tumors	Brain_Tumor1	Brain_Tumor2
SVM	OVR	**93.27**%	67.06%	94.99%	75.29%	**90**%	75.5%
	OVO	**93.27**%	54.63%	90.22%	46.39%	**90**%	73.83%
	DAG	**93.27**%	54.63%	90.22%	45.10%	**90**%	73.83%
	WW	**93.27**%	68.17%	94.31%	65.84%	**90**%	**77.17**%
	CS	**93.27**%	68.17%	94.31%	**75.38**%	**90**%	75.5%
SR		92.27%	**68.79**%	**95.02**%	74.04%	**90**%	**80.83**%

Table 2: Results with gene selection.

Method	Prostate_Tumor	9_Tumors	11_Tumors	14_Tumors	Brain_Tumor1	Brain_Tumor2
SVM	**94.36**%	**72.89**%	**96.66**%	75.38%	**90**%	**82.83**%
	OVR	CS	OVR	CS	WW	OVR
	BW50	BW3000	KW1000	NG*	NG*	KW500
SR	94.18%	72.40%	96.10%	**76.69**%	**90**%	80.83%
	BW800	BW3000	KW2000	BW5000	**NG***	NG*

* NG: no gene selection.

(v) Brain_Tumor2 [33]: there are 4 types of malignant glioma in this dataset: classic glioblastomas, classic anaplastic oligodendrogliomas, nonclassic glioblastomas, and nonclassic anaplastic oligodendrogliomas. The dataset has 50 samples, and the number of genes is 10 367.

(vi) Prostate_Tumor [9]: the binary dataset contains gene expression data of prostate tumor and normal tissues. There are 10 509 genes in each sample and 102 samples.

According to [25], 9_Tumors, 14_tumors, and Brain_Tumor2 are the most difficult datasets which make all the classifiers, including SVMs, generate low classification performance.

All the gene expression data are normalized by being rescaled between 0 and 1. It is also for the purpose of speeding up the training of SVMs.

3. Results and Discussion

Table 1 shows the classification results of the experiment without gene selection for both sparse representation (SR) and SVMs. The results of SVMs are slightly differently from [25]. A possible explanation is that the distribution file of cross validation is different in our study from [25]. From Table 1, the proposed SR approach performs better than all SVM variants on 9_Tumors, 11_Tumors, and Brian_Tumor2, and most SVM variants on 14_Tumors, while the SR approach performs comparably with SVM variants on Prostate_Tumor and Brain_Tumor1. In addition, similar to SVMs, the SR approach also finds it difficult to classify three multicategory datasets: 9_Tumors, 14_Tumors, and Brain_Tumor2. However, the SR approach performs better than all SVM variants on these datasets except CS and OVR on 14_Tumors. The difficulty may mainly be caused by the small number of total samples and even the smaller number of samples for each category. For example, the 9_Tumors dataset only has 60 samples, and category 7 (prostate tumor) just has two samples.

Table 2 shows the results of sparse representation when KW and BW methods are used for gene selection, along with the best results achieved by SVMs with the corresponding gene selection methods. From Table 2, the performance of the proposed SR is comparable with the best SVM variant on all six datasets. In addition, since gene selection generate limited improvement for both methods, sparse representation approach, similar to SVMs, seems less sensitive to curse of dimensionality than non-SVM methods such as neural network and k-nearest neighbors.

It is worth mentioning that the results of SVMs for both with and without gene selection are obtained by careful model selection using 9-fold cross validation. Spare representation approach, on the other hand, has no need of adjusting model parameters for different datasets.

As for the computing efficiency, sparse representation approach is very fast when sample number is less than 100. For example, without gene selection, it needs less than 10 seconds for Brain_Tumor2 dataset, which has only 50 samples. The efficiency, however, is dramatically reduced for relatively large sample cases. The dataset 14_Tumors, which has 308 samples, needs more than 3000 seconds! The main reason lies in the fact that the current implementation needs solving one optimization problem defined in (5) for classification of each testing sample. As a result, the number of optimization problems to be solved equals to the number of samples in the dataset. When compared with SVMs, however, the proposed SR is still faster, at least, than GEMS implementations when model selection is counted for SVMs.

4. Conclusion

In this paper, we have described a new approach for cancer diagnosis using gene expression data. The new method expresses each testing sample as a linear combination of all the training samples. The coefficient vector is obtained by l_1-regularized least square. Classification is achieved by defining discriminating functions from the coefficient vector for each category. Since l_1-norm minimization leads to sparse solution, we call the new approach sparse representation.

Numerical experiments show that sparse representation approach can match the best performance achieved by SVMs. Furthermore, the new approach has no need of model selection. One direction of our future work is to investigate how to classify multiple testing samples by solving only one optimization problem to improve the efficiency.

Acknowledgments

The second author would like to thank Natural Science and Engineering Research Council of Canada (NSERC) for supporting this research. Both authors thank the editor and reviewers for their kind comments and suggestions.

References

[1] R. Rifkin, S. Mukherjee, P. Tamayo, et al., "An analytical method for multiclass molecular cancer classification," *SIAM Review*, vol. 45, no. 4, pp. 706–723, 2003.

[2] T. R. Golub, D. K. Slonim, P. Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[3] L. J. van't Veer, H. Dai, M. J. van de Vijver, et al., "Expression profiling predicts outcome in breast cancer," *Breast Cancer Research*, vol. 5, no. 1, pp. 57–58, 2003.

[4] M. A. Shipp, K. N. Ross, P. Tamayo, et al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[5] A. Bhattacharjee, W. G. Richards, J. Staunton, et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13790–13795, 2001.

[6] L. Dyrskjøt, T. Thykjaer, M. Kruhøffer, et al., "Identifying distinct classes of bladder carcinoma using microarrays," *Nature Genetics*, vol. 33, no. 1, pp. 90–96, 2002.

[7] F. Bertucci, S. Salas, S. Eysteries, et al., "Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters," *Oncogene*, vol. 23, no. 7, pp. 1377–1391, 2004.

[8] G. P. Sawiris, C. A. Sherman-Baust, K. G. Becker, C. Cheadle, D. Teichberg, and P. J. Morin, "Development of a highly specialized cDNA array for the study and diagnosis of epithelial ovarian cancer," *Cancer Research*, vol. 62, no. 10, pp. 2923–2928, 2002.

[9] D. Singh, P. G. Febbo, K. Ross, et al., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[10] N. H. Segal, P. Pavlidis, W. S. Noble, et al., "Classification of clear-cell sarcoma as a subtype of melanoma by genomic profiling," *Journal of Clinical Oncology*, vol. 21, no. 9, pp. 1775–1781, 2003.

[11] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, et al., "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[12] D. Chen, Z. Liu, X. Ma, and D. Hua, "Selecting genes by test statistics," *Journal of Biomedicine and Biotechnology*, vol. 2005, no. 2, pp. 132–138, 2005.

[13] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[14] J. W. Lee, J. B. Lee, M. Park, and S. H. Song, "An extensive comparison of recent classification tools applied to microarray data," *Computational Statistics & Data Analysis*, vol. 48, no. 4, pp. 869–885, 2005.

[15] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[16] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[17] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[18] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[19] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[20] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale l_1-regularized least squares," *IEEE Journal on Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 606–617, 2007.

[21] E. Candès and J. Romberg, "l_1—magic: a collection of MATLAB routines for solving the convex optimization programs central to compressive sampling," 2006, http://www.acm.caltech.edu/l1magic.

[22] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[23] A. I. Su, J. B. Welsh, L. M. Sapinoso, et al., "Molecular classification of human carcinomas by use of gene expression signatures," *Cancer Research*, vol. 61, no. 20, pp. 7388–7393, 2001.

[24] S. Ramaswamy, P. Tamayo, R. Rifkin, et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 15149–15154, 2001.

[25] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.

[26] U. Kressel, "Pairwise classification and support vector machines," in *Advances in Kernel Methods: Support Vector Learning*, chapter 15, pp. 255–268, MIT Press, Cambridge, Mass, USA, 1999.

[27] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAGs for multiclass classification," in *Advances in Neural Information Processing Systems 12*, pp. 547–553, MIT Press, Cambridge, Mass, USA, 2000.

[28] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition," in *Proceedings of the 7th European Symposium on Artificial Neural Networks (ESANN '99)*, pp. 219–224, Bruges, Belgium, April 1999.

[29] K. Crammer and Y. Singer, “On the learnability and design of output codes for multiclass problems,” in *Proceedings of the 13th Annual Conference on Computational Learning Theory (COLT ’00)*, pp. 35–46, Standford University, Palo Alto, Calif, USA, June-July 2000.

[30] J. D. Gibbons, *Nonparametric Statistical Inference*, CRC Press, Boca Raton, Fla, USA, 4th edition, 2003.

[31] S. Dudoit, J. Fridlyand, and T. P. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data,” *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.

[32] J. E. Staunton, D. K. Slonim, H. A. Coller, et al., “Chemosensitivity prediction by transcriptional profiling,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 19, pp. 10787–10792, 2001.

[33] C. L. Nutt, D. R. Mani, R. A. Betensky, et al., “Gene expression-based classification of malignant gliomas correlates better with survival than histological classification,” *Cancer Research*, vol. 63, no. 7, pp. 1602–1607, 2003.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 617946, 8 pages
doi:10.1155/2009/617946

Methodology Report

Regularized F-Measure Maximization for Feature Selection and Classification

Zhenqiu Liu,[1] Ming Tan,[1] and Feng Jiang[2]

[1] *Division of Biostatistics, University of Maryland Greenebaum Cancer Center, 22 South Greene Street, Baltimore, MD 21201, USA*
[2] *Department of Pathology, University of Maryland at Baltimore, Baltimore, MD 21201, USA*

Correspondence should be addressed to Zhenqiu Liu, zliu@umm.edu

Received 22 December 2008; Accepted 17 March 2009

Recommended by Dechang Chen

Receiver Operating Characteristic (ROC) analysis is a common tool for assessing the performance of various classifications. It gained much popularity in medical and other fields including biological markers and, diagnostic test. This is particularly due to the fact that in real-world problems misclassification costs are not known, and thus, ROC curve and related utility functions such as F-measure can be more meaningful performance measures. F-measure combines recall and precision into a global measure. In this paper, we propose a novel method through regularized F-measure maximization. The proposed method assigns different costs to positive and negative samples and does simultaneous feature selection and prediction with L_1 penalty. This method is useful especially when data set is highly unbalanced, or the labels for negative (positive) samples are missing. Our experiments with the benchmark, methylation, and high dimensional microarray data show that the performance of proposed algorithm is better or equivalent compared with the other popular classifiers in limited experiments.

1. Introduction

Receiver Operating Characteristic (ROC) analysis has received increasing attention in the recent statistics and machine learning literatures (Pepe [1, 2]; Pepe and Janes [3]; Provost and Fawcett [4]; Lasko et al. [5]; Kun et al. [6]). ROC analysis originates in signal detection theory and is widely used in medical statistics for visualization and comparison of performance of binary classifiers. Traditionally, evaluation of a classifier is done by minimizing an estimation of a generalization error or some other related measures (Vapnik [7]). However the accuracy (the rate of correct classification) of a model does not always work. In fact when the data are highly unbalanced, accuracy may be misleading, since the all-positive or all-negative classifiers may achieve very good classification rate. In real life applications, the situations for which the data sets are unbalanced arise frequently. Utility functions such as F-measure or AUC provide a better way for classifier evaluation, since they can assign different error costs for positive and negative samples.

When the goal is to achieve the best performance under a ROC-based utility functions, it may be better to build classifiers through directly optimizing the utility functions. In fact, optimizing the log-likelihood function or the mean-square error does not necessarily imply good ROC curve performance. Hence, several algorithms have been recently developed for optimizing the area under ROC curve (AUC) function (Freund et al. [8]; Cortes and Mohri [9]; Rakotomamonjy [10]), and they have been proven to work well with different degrees of success. However, there are not many methods proposed for F-measure maximization. Most approaches to date that we know of maximize F-measure using SVMs and do so by varying parameters in standard SVM in an attempt to maximize F-measure as much as possible (Musicant et al. [11]). While this may result in a "best possible" F-measure for a standard SVM, there is no evidence that this technique should produce an F-measure comparable with one from the classifier designed to specifically optimize F-measure. Jansche [12] proposed an approximation algorithm for F-measure maximization in the logistic regression framework. His method, however, gives extremely large values for the estimated parameters and creates too many steep gradients. It, therefore, either converges very slow or fails to converge for large datasets.

Table 1: Classification outcomes.

		Predicted		Total
		1	−1	
True	1	TP	FN	N_p
	−1	FP	TN	N_n
		M_p	M_n	

Our aim in this paper is to propose a novel algorithm that directly optimizes an approximation of the regularized F-measure. The regularization term can be an L_2, L_1 or a combination of L_1 and L_2 penalty based on different prior assumptions (Tibshirani [13, 14]; Wang et al. [15]). Due to the nature of L_1 penalty, our algorithm provides simultaneous feature selection and classification with L_1 penalty. The proposed algorithm can be easily applied to high dimensional microarray data. One advantage with this method is that it is very efficient when data is highly unbalanced, since it assigns different costs to the positive and negative samples.

The paper is organized as follows. In Section 2 we introduce the related concept of ROC and F-measure. The algorithm and the brief proof of its generalization bounds are proposed in Section 3. The computational experiments and performance evaluation are given in Section 4. Finally the conclusions and remarks are discussed in Section 5.

2. ROC Curves and F-Measure

In binary classification, a classifier attempts to map the instances into two classes: positive (p) and negative (n). There are four possible outcomes with the given classifier: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Table 1 summarizes these outcomes with their associated terminology. The number of positive instances is $N_\mathrm{p} = \mathrm{TP} + \mathrm{FN}$. Similarly $N_\mathrm{n} = \mathrm{TN} + \mathrm{FP}$ is the number of negative instances.

From these counts the following statistics are derived:

$$\begin{aligned} \mathrm{tpr} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \qquad \mathrm{tnr} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \\ \mathrm{fpr} &= \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}, \qquad \mathrm{fnr} = \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}}, \end{aligned} \tag{1}$$

where true positive rate (also called recall or sensitivity) is denoted by tpr and true negative rate (specificity) by tnr. False positive rate and false negative rate are denoted by fpr and fnr, respectively. Note that $\mathrm{tnr} = 1 - \mathrm{fpr}$, and $\mathrm{fnr} = 1 - \mathrm{tpr}$. We also define the precision $\mathrm{Pr} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$. ROC curves plot the true positive rate versus false positive rate by varying the threshold which is usually the probability of the membership to a class, distance to a decision surface, or a score produced by a decision function. In the ROC space, the upper left corner represents a perfect classification, while a diagonal line represents random classification. A point in ROC curve that lies upper left of another point represents a better model.

F-measure combines the true positive rate (recall) and precision Pr into a single utility function which is defined as γ-weighted harmonic mean:

$$F_\gamma = \frac{1}{\gamma(1/\mathrm{tpr}) + (1-\gamma)(1/\mathrm{Pr})}, \quad \text{where } 0 \le \gamma \le 1. \tag{2}$$

F_γ can be expressed with TP, FP, and FN as follows:

$$F_\gamma = \frac{\mathrm{TP}}{\mathrm{TP} + \gamma\mathrm{FN} + (1-\gamma)\mathrm{FP}} \tag{3}$$

or equivalently

$$F_\gamma = \frac{\mathrm{TP}}{\gamma N_\mathrm{p} + (1-\gamma)M_\mathrm{p}}, \tag{4}$$

where N_p is the number of positive samples, and $M_\mathrm{p} = \mathrm{TP} + \mathrm{FP}$. Clearly $0 \le F_\gamma \le 1$ and $F_\gamma = 1$ only when all the data are classified correctly. Maximizing F-measure is equivalent to maximizing the weighted sensitivity and specificity. Therefore, maximizing F_γ will indirectly lead to maximize the area under ROC curve (AUC).

To optimize F_γ, we have to define TP, FN, and FP mathematically. We first introduce an indicator function

$$I(y \in C) = \begin{cases} 1, & \text{if } y \in C, \\ 0, & \text{if } y \notin C, \end{cases} \tag{5}$$

where C is a set. Let $y = f(\mathbf{w}, \mathbf{x})$ be a classifier with coefficients (weights) $\mathbf{w}$ and input variable $\mathbf{x}$, and let $\hat{y}$ be the predicted value. Given n samples, $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i$ is a multidimensional input vector with dimension m and class label $y_i \in \{-1, 1\}$; TP, FN, and FP are given, respectively:

$$\begin{aligned} \mathrm{TP} &= \sum_{i=1}^{n} I(\hat{y}_i = 1) I(y_i = 1), \\ \mathrm{FN} &= \sum_{i=1}^{n} I(\hat{y}_i = -1) I(y_i = 1), \end{aligned} \tag{6}$$

$$\mathrm{FP} = \sum_{i=1}^{n} I(\hat{y}_i = 1) I(y_i = -1). \tag{7}$$

It is clear that F-measure is a utility function that applies for the whole data set.

3. The Algorithm

Usually given a classifier with known parameters $\mathbf{w}$, F-measure can be calculated with the test data to evaluate the performance of the model. The aim of this paper is, however, to learn a classifier and estimate the corresponding parameters $\mathbf{w}$ with a given training data D and regularized F-measure maximization. Since $F_\gamma \in [0, 1]$, we have $-\log F_\gamma \in [0, \infty)$. Statistically F_γ is a probability that measures the proportion of samples correctly classified. Based on these observations, we can maximize the $\log F_\gamma$ in the maximum log likelihood framework. Different assumptions for the prior distribution of $\mathbf{w}$ will lead to different penalty terms. Given the coefficient vector $\mathbf{w}$ with dimension m, we have $L_2 = (1/2)\sum_{j=1}^{m}|w_j|^2$ for the assumption of Gaussian distribution and $L_1 = \sum_{j=1}^{m}|w_j|$ with that of Laplacian prior. In general, L_1 penalty encourages sparse solutions, while the classifiers with L_2 are more robust. We make TP, FN, and FP depend on $\mathbf{w}$ explicitly and maximize the following penalized F-measure functions:

$$\begin{aligned} E_1(\mathbf{w}) &= \log F_\gamma(\mathrm{TP}(\mathbf{w}), \mathrm{FN}(\mathbf{w}), \mathrm{FP}(\mathbf{w})) - \lambda \sum_{j=1}^{m}|w_j|, \\ E_2(\mathbf{w}) &= \log F_\gamma(\mathrm{TP}(\mathbf{w}), \mathrm{FN}(\mathbf{w}), \mathrm{FP}(\mathbf{w})) - \frac{1}{2}\lambda \sum_{j=1}^{m}|w_j|^2. \end{aligned} \tag{8}$$

We have

$$\begin{aligned} \hat{\mathbf{w}} &= \arg\max_{\hat{\mathbf{w}}}\left\{ \log F_\gamma(\mathrm{TP}(\mathbf{w}), \mathrm{FN}(\mathbf{w}), \mathrm{FP}(\mathbf{w})) - \lambda \sum_{j=1}^{m}|w_j| \right\}, \\ \hat{\mathbf{w}} &= \arg\max_{\hat{\mathbf{w}}}\left\{ \log F_\gamma(\mathrm{TP}(\mathbf{w}), \mathrm{FN}(\mathbf{w}), \mathrm{FP}(\mathbf{w})) - \frac{1}{2}\lambda \sum_{j=1}^{m}|w_j|^2 \right\}. \end{aligned} \tag{9}$$

Note that $\mathrm{TP}(\mathbf{w})$, $\mathrm{FN}(\mathbf{w})$, and $\mathrm{FP}(\mathbf{w})$ are all integers, and the index function I in (7) is not differentiable. We first define an S-type function to approximate the index function I: Let $z = \mathbf{w}^T\mathbf{x}$ be a linear score function,

$$h(z) = \begin{cases} 0, & z < -1, \\ \frac{1}{2}(1+z)^2, & -1 \le z \le 0, \\ \frac{1}{2}(2-(1-z)^2), & 0 < z \le 1, \\ 1, & z > 1. \end{cases} \tag{10}$$

The decision role such that $\hat{y}(\mathbf{w}, \mathbf{x}) = 1$ if $z = \mathbf{w}^T\mathbf{x} > 0$ can be represented as

$$I(\hat{y} = 1) = I(z > 0) = I(h(\mathbf{w}^T\mathbf{x}) > 0.5) \approx h(\mathbf{w}^T\mathbf{x}). \tag{11}$$

Figure 1 gives some insight about the $h(z)$. Figure 1 shows that $h(z)$ is a better approximation of $I(z > 0)$ than the sigmoid function $g(z) = 1/(1+e^{-z})$. The first derivative of $h(z)$ is continuous and given in (12):

$$h'(z) = \frac{dh(z)}{dz} = \begin{cases} 0, & z < -1, \\ 1+z, & -1 \le z \le 0, \\ 1-z, & 0 < z \le 1, \\ 0, & z > 1. \end{cases} \tag{12}$$

Based on (10) and (11), the approximated version of $\mathrm{TP}(\mathbf{w})$ and $M_\mathrm{p}(\mathbf{w}) = \mathrm{TP}(\mathbf{w}) + \mathrm{FP}(\mathbf{w})$ can be written as follows:

$$\begin{aligned} \mathrm{TP}(\mathbf{w}) &= \sum_{\substack{i=1 \\ y_i=1}}^{n} h(\mathbf{w}^T\mathbf{x}_i), \\ M_\mathrm{p}(\mathbf{w}) &= \sum_{i=1}^{n} h(\mathbf{w}^T\mathbf{x}_i). \end{aligned} \tag{13}$$

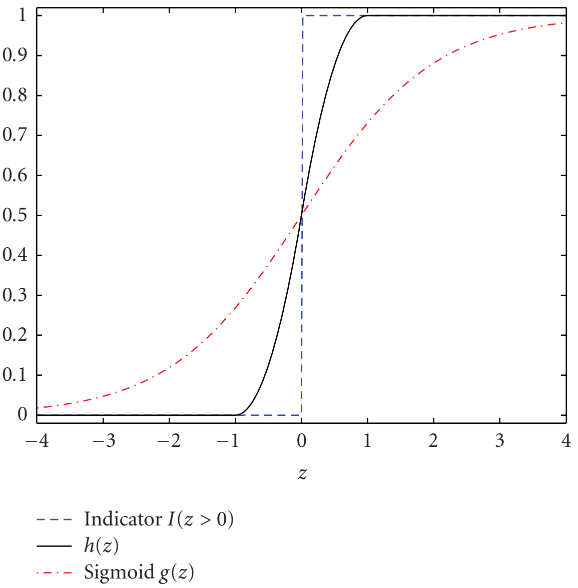


Figure 1: The plot for $h(z)$, indicator function $I(z > 0)$, and Sigmoid $g(z) = 1/(1+e^{-z})$.

We can find the first-order derivatives of E_1 and E_2, respectively, as follows:

$$\begin{aligned} \frac{\partial E_1(\mathbf{w})}{\partial w_j} &= \frac{\partial F_\gamma(\mathbf{w})/\partial w_j}{F_\gamma(\mathbf{w})} - \lambda\,\mathrm{sign}\,(w_j), \\ \frac{\partial E_2(\mathbf{w})}{\partial w_j} &= \frac{\partial F_\gamma(\mathbf{w})/\partial w_j}{F_\gamma(\mathbf{w})} - \lambda w_j, \end{aligned} \tag{14}$$

where,

$$\begin{aligned} \frac{\partial F_\gamma(\mathbf{w})}{\partial w_j} &= B\frac{\partial \mathrm{TP}(\mathbf{w})}{\partial w_j} - B^2\mathrm{TP}(\mathbf{w})(1-\gamma)\frac{\partial M_\mathrm{p}(\mathbf{w})}{\partial w_j}, \\ B &= \frac{1}{\gamma N_p + (1-\gamma)M_\mathrm{p}(\mathbf{w})}, \\ \frac{\partial \mathrm{TP}(\mathbf{w})}{\partial w_j} &= \sum_{\substack{i=1 \\ y_i=1}}^{n} h'(\mathbf{w}^T\mathbf{x})x_{ij}, \\ \frac{\partial M_\mathrm{p}(\mathbf{w})}{\partial w_j} &= \sum_{i=1}^{n} h'(\mathbf{w}^T\mathbf{x})x_{ij}. \end{aligned} \tag{15}$$

Knowing E_1 and E_2, and their derivatives $\nabla E_1 = [\partial E_1/\partial w_j]$ and $\nabla E_2 = [\partial E_2/\partial w_j]$, we can maximize the penalized function E_1 and E_2 with gradient descent-related algorithm such as Broyden-Fletcher-Goldfarb-Shanno- (BFGS-) related quasi-Newton method (Broyden [16]). The algorithm for E_2 maximization is straight forward as shown in Algorithm 1. The step-size μ in the algorithm can be found with line search.

The regularized F-measure maximization with L_1 penalty (E_1) is of especial interest because it favors sparse solutions and can select features automatically. However, maximizing E_1 is a little bit complex since L_1 and E_1 are not differentiable at 0. For simplicity, let $LF = \log F_\gamma(\mathbf{w})$, we have

```
1. Given $\gamma$, $\lambda$, a small number $\varepsilon$, Initialize $\mathbf{w}^t = \mathbf{w}^0$, and set
   $t = 0$.
2. While $|\mathbf{w}^{t+1} - \mathbf{w}^t| > \varepsilon$
   $\mathbf{w}^{t+1} = \mathbf{w}^t + \mu(\nabla E_2)$, where $\mu$ is the step-size
3. $t = t + 1$
```

Algorithm 1: L_2 regularized F-measure maximization.

```
1. Given $\gamma$, $\lambda$, small numbers $\varepsilon$ and $\delta$, $\mathbf{w}^t = \mathbf{w}^0$, and set
   $t = 0$ and $\Psi = \{j : w_j \neq 0\}$.
2. While $|\mathbf{w}^{t+1} - \mathbf{w}^t| > \varepsilon$
   $\mathbf{w}^{t+1} = \mathbf{w}^t + \mu\left(\left(\frac{\partial LF}{\partial w_j}\right)_\Psi - \lambda\,\mathrm{sign}\,(w_i)_\Psi\right)$, where $\mu$ is
   the step-size
   $\Psi = \Psi \cup \left\{ j \notin \Psi : \left|\frac{\partial LF}{w_j}\right| > \lambda \right\}$
   $\Psi = \Psi \setminus \{ j \in \Psi : |w_j| < \delta \}$
3. $t = t + 1$
```

Algorithm 2: L_1 regularized F-measure maximization.

$E_1 = LF - \lambda\sum_{j=1}^{m}|w_j|$. The Karush-Kuhn-Tucker (KKT) conditions for optimality are given as follows:

$$\left|\frac{\partial LF}{\partial w_j}\right| < \lambda \Longrightarrow w_j = 0,$$
$$w_j \neq 0 \Longrightarrow \left|\frac{\partial LF}{\partial w_j}\right| = \lambda. \tag{16}$$

The KKT conditions tell us that we have a set Ψ of nonzero coefficients which corresponds to the variables whose absolute value of first-order derivative is maximal and equal to λ, and that all variables with smaller derivatives have zero coefficients at the optimal penalized solution. Since L_1 is differentiable everywhere except at 0, we can design an algorithm to deal with the nonzero coefficients only. Algorithm 2 proposes an algorithm that can be applied to the subspace of nonzero coefficient set denoted by Ψ. The algorithm has a procedure to add or remove variables from Ψ, when the first-order derivative becomes large and when a coefficient hits 0, respectively.

3.1. Computational Considerations. Both γ and λ are free parameters that need to be chosen. We will choose the best parameter for γ and λ with the area under ROC curve (AUC). Area under the ROC curve (AUC) is another scalar measure for classifier comparison. Its value is between $(0, 1)$. Larger AUC values indicate better classifier performance across the full range of possible thresholds. For datasets with skewed class or cost distribution is unknown as in our applications, AUC is a better measure than prediction accuracy.

Given a binary classification problem with N_p positive class samples and N_n negative class samples, let $f(\mathbf{x})$ be the score function to rank a sample $\mathbf{x}$. AUC is the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Mathematically

$$\mathrm{AUC} = \frac{\sum_{i=1}^{N_p}\sum_{j=1}^{N_n} I(f(\mathbf{x}_i) > f(\mathbf{y}_j))}{N_p N_n}, \tag{17}$$

where $I(\cdot)$ is an index function and $I(\cdot) = 1$ if $f(\mathbf{x}_i) > f(\mathbf{y}_j)$, otherwise $I(\cdot) = 0$. AUC is also called Wilcoxon-Mann-Whitney statistic (Rakotomamonjy [10]).

Table 2: Overview of the datasets.

Datasets	No. of samples (train/test)	No. of variables	No. of experiments
Breast cancer	200/77	9	100
Diabetis	468/300	8	100
Heart	170/100	13	100
German	700/300	20	100
Thyroid	140/75	5	100
Titanic	150/2051	3	100

Note that $\log F_\gamma(\mathbf{w})$ is generally a nonconcave function with respect to $\mathbf{w}$; only local maximum is guaranteed. One way to deal with this difficulty is to employ the multiple-points initialization. Multiple random points are generated, and our proposed algorithms are used to find the maximum for each point. The result with the lowest test error is chosen as our best solution.

4. Computational Results

4.1. Benchmark Data. To evaluate the performance of the proposed method, experiments were performed on six benchmark datasets which can be downloaded from http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm. These benchmark datasets have been widely used in model comparison studies in machine learning. They are all binary classification problems, and the datasets were randomly divided into train and test data 100 times to prevent bias and overfitting. The data are normalized with zero mean and standard deviation. The overview of the datasets is given in Table 2. The computational results with our algorithms, logistic regression, and linear support vector machines are given in Figures 2-3.

Figures 2-3 show that L_2 F-measure maximization performs better or equivalent compared with logistic regression and linear support vector machines (SVM) in limited experiments. In fact, the test errors for all datasets except for Thyroid are competitive with that of the nonlinear classification methods reported by Ratsch (http://ida.first.fraunhofer.de/projects/bench/benchmarks .htm). The inferior performance of L_2 F-measure with Thyroid data indicates the strong nonlinear factors in that data.

4.2. Real Methylation Data. This methylation data are from 7 CpG regions and 87 lung cancer cell lines (Virmani et al. [17], Siegmund et al. [18]). 41 lines are

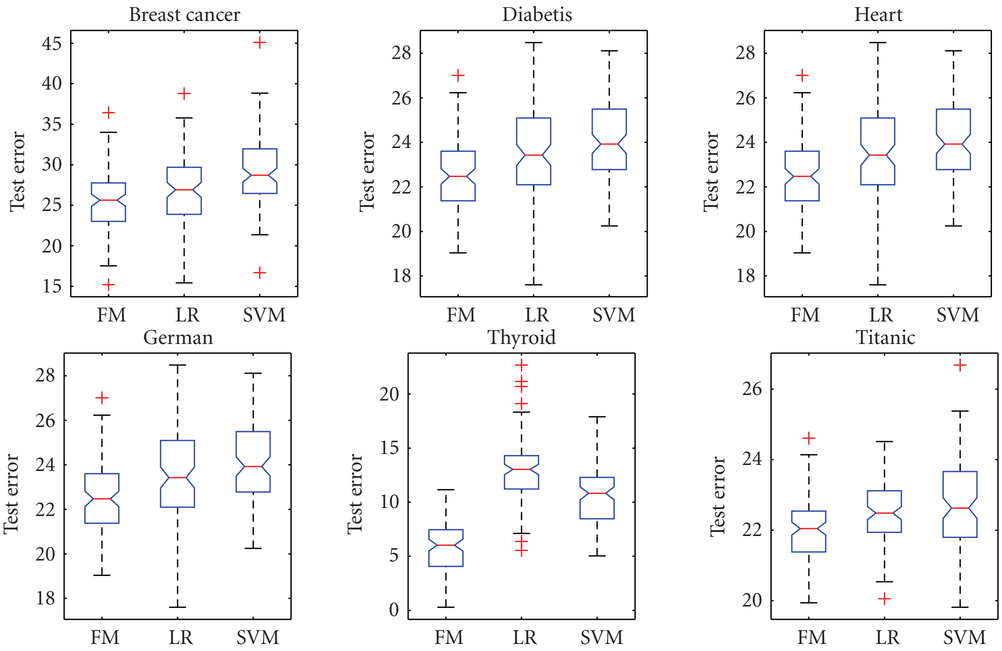


Figure 2: Test Errors of L_2 F-Measure (FM), Logistic Regression (LR), and SVM.

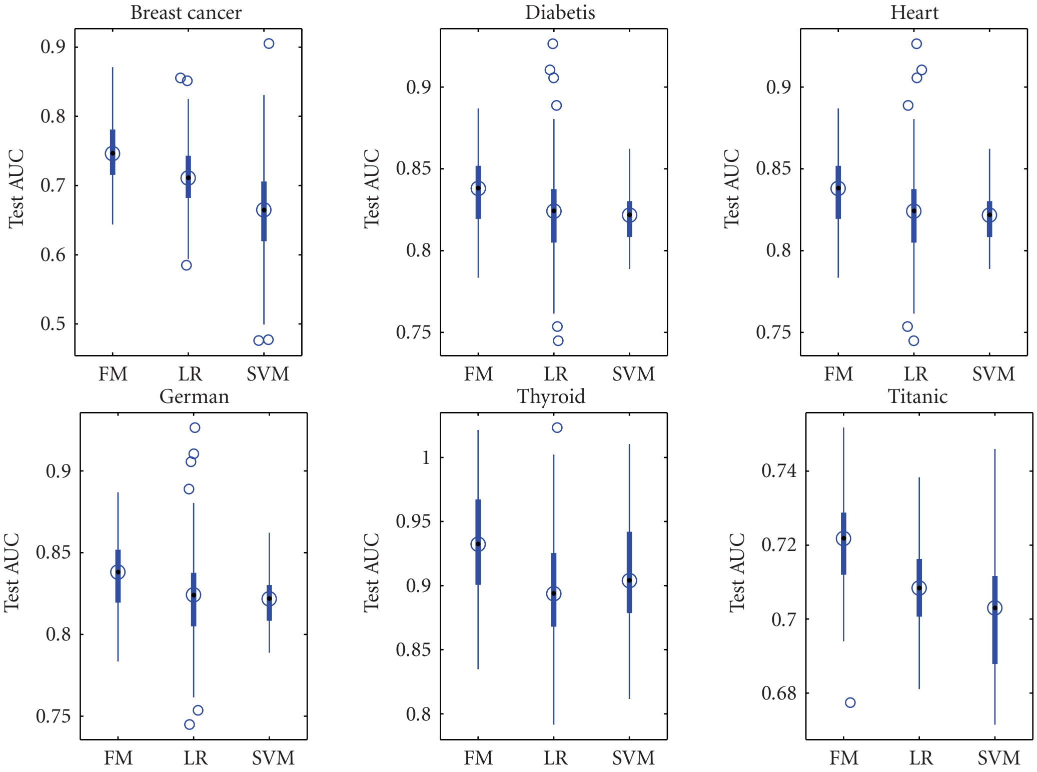


Figure 3: Test AUC of L_2 F-Measure (FM), Logistic Regression (LR), and SVM.

from small cell lung cancer and 46 lines from nonsmall cell lung cancer. The proportion of positive values for the different regions ranges from 39% to 100% for the small cell lung cancer and from 65% to 98% for the nonsmall cell lung cancer. The data are available at http://www-rcf.usc.edu/kims/SupplementaryInfo.html. We utilize the twofold cross validation scheme to choose the best λ and test our algorithms. Other cross-validation schemes such as 10-fold cross validation will lead to similar results but are more computational intensive. We randomly split

Table 3: Performance with different γ's and L_1 F-measure maximization.

γ	Variables selected (1/0)	Sensitivity	Specificity	Test error	AUC
0.1	1101111	0.476	0.957	27.3	0.801
0.2	1111111	0.714	0.957	**15.9**	0.820
0.3	0101111	0.810	0.740	22.7	0.849
0.4	1111001	0.826	0.762	20.4	**0.861**
0.5	1111101	0.857	0.609	27.3	0.832
0.6	1100101	0.762	0.739	25	0.832
0.7	1110110	0.904	0.348	38	0.847
0.8	1011100	100	0.217	40.9	0.826
0.9	1100011	100	0	52.3	0.754

the data into two roughly equal-sized subsets and build the classifier with one subset and test it with the other. To avoid the bias arising from a particular partition, the procedure is repeated 100 times, each time splitting the data randomly into two folds and doing the cross validation. The average computational results with different γs and $\lambda = 0.05$ are given in Table 3. Table 3 shows the selected variables (1: selected; 0: not selected), sensitivity, specificity, test errors, and AUC values with different γ's. We can see clearly the sensitivity increases while the specificity decreases as γ increases. When $\gamma = 0.9$, every example is classified as positive examples. The best γ will be 0.4 according to AUC but it will be 0.2 based on test error. Therefore, again there is some inconsistence between two measures. Figure 4 gives some sight about how to choose λ and the number of features. Given $\gamma = 0.4$, the optimal $\lambda = 0.04$, and those 5 out of 7 CpG regions selected by L_1 F-measure maximization have been proved to be predictive of lung cancer subtype (Siegmund et al. [18]). The performance of the model is improved roughly 6% in AUC and 3% in test error with only 5 instead of 7 CpG regions.

4.3. High Dimensional Microarray Data. The colon microarray data set (Alon et al. [19]) has 2000 features (genes) per sample and 62 samples which consisted 22 normal and 40 cancer tissues. The task is to distinguish tumor from normal tissues. The data set was first normalized for each gene to have zero mean and unit variance. The transformed data was then used for all the experiments. We employed a same twofold cross validation scheme to evaluate the model. This computational experiments are repeated 100 times. The AUC was calculated after each cross validation. The computational results for performance comparison are reported in Table 4.

Table 4 gives us some insight that how the model performance changes with different γ's. Generally we can see that the false negative (FN) decreases and the false positive (FP) increases as γ increases. The only exception is when $\gamma = 0.1$, both FN and FP have the worst performance. The best performance is achieved when $\gamma \in [0.7, 0.8]$ according to both AUC and the number of misclassified samples.

Table 4: Performance with different γ's and L_1 F-measure maximization ($\lambda = 3$).

γ	No. of variables	FN	FP	No. of misclassified	AUC
0.1	10	11	33	44	0.588
0.2	10	3	3	6	0.989
0.3	10	3	3	6	0.989
0.4	10	3	3	6	0.989
0.5	10	3	3	6	0.989
0.6	10	3	3	6	0.989
0.7	10	2	3	5	**0.993**
0.8	10	2	3	5	**0.993**
0.9	10	2	5	7	0.988
1	10	2	8	10	0.971

Table 5: 10 differentially expressed genes.

Gene ID	Description
h20709	myosin light chain alkali, smooth-muscle isoform (human)
t71025	84103 human (human)
m76378	human cysteine-rich protein (crp) gene, exons 5 and 6
m63391	human desmin gene, complete cds
z50753	h.sapiens mrna for gcap-ii/uroguanylin precursor
r87126	myosin heavy chain, nonmuscle (gallus gallus)
x12671	human gene for heterogeneous nuclear ribonucleoprotein (hnrnp) core protein a1
t92451	tropomyosin, fibroblast and epithelial muscle-type (human)
j02854	myosin regulatory light chain 2, smooth muscle isoform (human); contains element tar1 repetitive element
m36634	human vasoactive intestinal peptide (vip) mrna, complete cds

The 10 genes selected are given in Table 5. The selected genes allow the separation of cancer from normal samples in the gene expression map. Some genes were selected because their activities resulted in the difference in the tissue composition between normal and cancer tissue. Other genes were selected because they played a role in cancer formation or cell proliferation. It was not surprise that some genes implicated in other types of cancer such as breast and prostate cancers were identified in the context

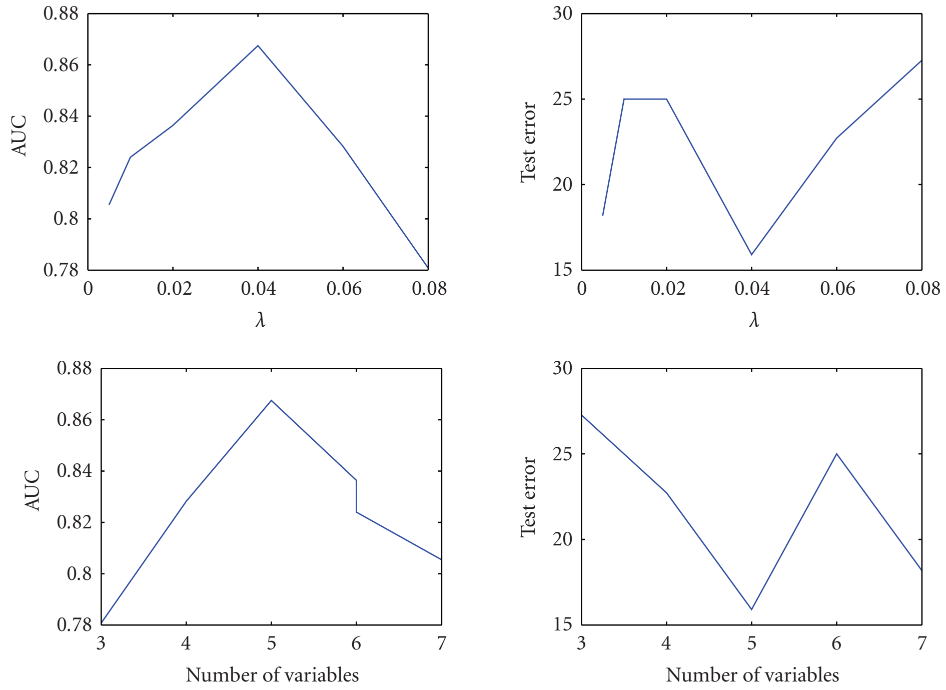


Figure 4: Performance with different λs and number of variables.

of colon cancer because these tissue types shared similarity. Our method is supported by the meaningful biological interpretation of selected genes. For instance, three muscle-related genes (H20709, T92451, and J02854) were selected from the colon cancer data, reflecting the fact that normal colon tissue had higher muscle content, whereas colon cancer tissue had lower muscle content (biased toward epithelial cells), and the selection of x12671 ribosomal protein agreed with an observation that ribosomal protein genes had lower expression in normal than in cancer colon tissue.

5. Conclusions and Remarks

We have presented a novel regularized F-measure maximization for feature selection and classification. This technique directly maximizes the tradeoff between specificity and sensitivity. Regularization with L_2 and L_1 allows the algorithm to converge quickly and to do simultaneous feature selection and classification. We found that it has better or equivalent performances when compared with the other popular classifiers in limited experiments.

The proposed method has the ability to incorporate nonstandard tradeoffs between sensitivity and specificity with different γ. It is well suited for dealing with unbalanced data or data with missing negative (positive) samples. For instance, in the problem of gene function prediction, the available information is only about positive samples. In other words, we know which genes have the function of interested, while it is generally unclear which genes do not have the function. Most standard classification methods will fail but our method can train the model with only positive labels by setting $\gamma = 1$.

One difficulty with the regularized F-measure maximization is the nonconcavity of the error function. We utilized the random multiple points initialization to find the optimal solutions. More efficient algorithms for nonconcave optimization will be considered to speed up the computations. The applications of the proposed method in gene function predictions and others will be explored in the future.

References

[1] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, Oxford, UK, 2003.

[2] M. S. Pepe, "Evaluating technologies for classification and prediction in medicine," *Statistics in Medicine*, vol. 24, no. 24, pp. 3687–3696, 2005.

[3] M. S. Pepe and H. Janes, "Insights into latent class analysis of diagnostic test performance," *Biostatistics*, vol. 8, no. 2, pp. 474–484, 2007.

[4] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Machine Learning*, vol. 42, no. 3, pp. 203–231, 2001.

[5] T. A. Lasko, J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado, "The use of receiver operating characteristic curves in biomedical informatics," *Journal of Biomedical Informatics*, vol. 38, no. 5, pp. 404–415, 2005.

[6] D. Kun, C. Bourke, S. Scott, and N. V. Vinodchandran, "New algorithms for optimizing multi-class classifiers via ROC surfaces," in *Proceedings of the 3rd Workshop on ROC Analysis in Machine Learning (ROCML '06)*, pp. 17–24, Pittsburgh, Pa, USA, June 2006.

[7] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[8] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research*, vol. 4, no. 6, pp. 933–969, 2004.

[9] C. Cortes and M. Mohri, "AUC optimization vs. error rate minimization," in *Advances in Neural Information Processing Systems 16*, pp. 313–320, MIT Press, Cambridge, Mass, USA, 2003.

[10] A. Rakotomamonjy, "Optimizing AUC with Support Vector Machine (SVM)," in *Proceedings of European Conference on Artificial Intelligence Workshop on ROC Curve and AI*, Valencia, Spain, 2004.

[11] D. R. Musicant, V. Kumar, and A. Ozgur, "Optimizing F-measure with support vector machines," in *Proceedings of the 16th International Florida Artificial Intelligence Research Society Conference (FLAIRS '03)*, pp. 356–360, St. Augustine, Fla, USA, May 2003.

[12] M. Jansche, "Maximum expected F-measure training of logistic regression models," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP '05)*, pp. 692–699, Vancouver, Canada, October 2005.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[14] R. Tibshirani, "The lasso method for variable selection in the Cox model," *Statistics in Medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[15] L. Wang, J. Zhu, and H. Zou, "The doubly regularized support vector machine," *Statistica Sinica*, vol. 16, no. 2, pp. 589–615, 2006.

[16] C. G. Broyden, "Quasi-Newton methods and their application to function minimization," *Mathematics of Computation*, vol. 21, no. 99, pp. 368–381, 1967.

[17] A. K. Virmani, J. A. Tsou, K. D. Siegmund, et al., "Hierarchical clustering of lung cancer cell lines using DNA methylation markers," *Cancer Epidemiology Biomarkers & Prevention*, vol. 11, no. 3, pp. 291–297, 2002.

[18] K. D. Siegmund, P. W. Laird, and I. A. Laird-Offringa, "A comparison of cluster analysis methods using DNA methylation data," *Bioinformatics*, vol. 20, no. 12, pp. 1896–1904, 2004.

[19] U. Alon, N. Barkai, D. A. Notterman, et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 12, pp. 6745–6750, 1999.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 608701, 10 pages
doi:10.1155/2009/608701

Research Article

Gene-Based Multiclass Cancer Diagnosis with Class-Selective Rejections

Nisrine Jrad, Edith Grall-Maës, and Pierre Beauseroy

Institut Charles Delaunay (ICD, FRE CNRS 2848), Université de Technologie de Troyes, LM2S 12 rue Marie Curie, BP 2060, 10010 Troyes cedex, France

Correspondence should be addressed to Nisrine Jrad, nisrine.jrad@utt.fr

Received 15 January 2009; Accepted 13 March 2009

Recommended by Dechang Chen

Supervised learning of microarray data is receiving much attention in recent years. Multiclass cancer diagnosis, based on selected gene profiles, are used as adjunct of clinical diagnosis. However, supervised diagnosis may hinder patient care, add expense or confound a result. To avoid this misleading, a multiclass cancer diagnosis with class-selective rejection is proposed. It rejects some patients from one, some, or all classes in order to ensure a higher reliability while reducing time and expense costs. Moreover, this classifier takes into account asymmetric penalties dependant on each class and on each wrong or partially correct decision. It is based on ν-1-SVM coupled with its regularization path and minimizes a general loss function defined in the class-selective rejection scheme. The state of art multiclass algorithms can be considered as a particular case of the proposed algorithm where the number of decisions is given by the classes and the loss function is defined by the Bayesian risk. Two experiments are carried out in the Bayesian and the class selective rejection frameworks. Five genes selected datasets are used to assess the performance of the proposed method. Results are discussed and accuracies are compared with those computed by the Naive Bayes, Nearest Neighbor, Linear Perceptron, Multilayer Perceptron, and Support Vector Machines classifiers.

1. Introduction

Cancer diagnosis, based on gene expression profiling, have improved over the past 40 years. Many microarray technologies studies were developed to analyze the gene expression. These genes are later used to categorize cancer classes. Two different classification approaches can be used: class discovery and class prediction. The first is an unsupervised learning approach that allows to separate samples into clusters based on similarities in gene expression, without prior knowledge of sample identity. The second is a supervised approach which predicts the category of an already defined sample using its gene expression profiles. Since these classification problems are described by a large number of genes and a small number of samples, it is crucial to perform genes selection before the classification step. One way to identify informative genes pointed in [1] is the test statistics.

Researches show that the performance of supervised decisions based on selected gene expression can be comparable to the clinical decisions. However, no classification strategy is absolutely accurate. First, many factors may effectively decrease the predictive power of a multiclass problem. For example, findings of [2] imply that information useful for multiclass tumor classification is encoded in a complex gene expression and cannot be given by a simple one. Second, it is not possible to find an optimal classification method for all kinds of multiclass problems. Thus, supervised diagnosis are always considered as an important adjunct of traditional diagnostics and never like its substitute.

Unfortunately, supervised diagnosis can be misleading. They may hinder patient care (wrong decision on a sick patient), add expense (wrong decision on a healthy patient) or confound the results of cancer categories. To overcome these limitations, a multi-SVM [3] classifier with class-selective rejection [4–7] is proposed. Class-selective rejection consists of rejecting some patients from one, some, or all classes in order to ensure a higher reliability while reducing time and expense costs. Moreover, any of the existing multiclass [8–10] algorithms have taken into consideration asymmetric penalties on wrong decisions. For example, in a binary cancer problem, a wrong decision on a sick patient must cost more than a wrong decision on a healthy patient. The proposed classifier handles this kind of problems. It minimizes a general loss function that takes into account asymmetric penalties dependant on each class and on each wrong or partially correct decision.

The proposed method divides the multiple class problem into several unary classification problems and train one ν-1-SVM [11–13] coupled with its regularization path [14, 15] for each class. The winning class or subset of classes is determined using a prediction function that takes into consideration the costs asymmetry. The parameters of all the ν-1-SVMs are optimized jointly in order to minimize a loss function. Taking advantage of the regularization path method, the entire parameters searching space is considered. Since the searching space is widely extended, the selected decision rule is more likely to be the optimal one. The state-of-art multiclass algorithms [8–10] can be considered as a particular case of the proposed algorithm where the number of decisions is given by the existing classes and the loss function is defined by the Bayesian risk.

Two experiments are reported in order to assess the performance of the proposed approach. The first one considers the proposed algorithm in the Bayesian framework and uses the selected microarray genes to make results comparable with existing ones. Performances are compared with those assessed using Naive Bayes, Nearest Neighbor, Linear Perceptron, Multilayer Perceptron, and Support Vector Machines classifiers, invoked in [1]. The second one shows the ability of the proposed algorithm solving multiclass cancer diagnosis in the class-selective rejection scheme. It minimizes an asymmetric loss function. Experimental results show that, a cascade of class-selective classifiers with class-selective rejections can be considered as an improved supervised diagnosis rule.

This paper is outlined as follows. Section 2 presents a description of the model as a gene selection task. It introduces the multiclass cancer diagnosis problem in the class-selective rejection scheme. It also proposes a supervised training algorithm based on ν-1-SVM coupled with its regularization path. The two experiments are carried out in Section 3, results are reported, compared and discussed. Finally, a conclusion is presented in Section 4.

2. Models and Methods

This section describes the multiclass cancer diagnosis based on microarray data. Feature selection is evoked as a first process in a gene-based cancer diagnosis. Test statistics are used as a possible way for informative genes identification [1]. Once genes selection is processed, a classification problem should be solved. The multiclass cancer diagnosis problem, formulated in the general framework of class-selective rejection, is introduced. A solution based on ν-1-SVM [11–13] is proposed. First a brief description of ν-1-SVM and the derivation of its regularization path [14, 15] is presented. Second, the proposed algorithm [3] is explained. It allows to determine a multiclass cancer diagnosis that minimizes an asymmetric loss function in the class-selective rejection scheme.

2.1. Genes Selection Using Test Statistics. Gene profiles are successfully applied to supervised cancer diagnosis. Since cancer diagnosis problems are usually described by a small set of samples with a large number of genes, feature or gene selection is an important issue in analyzing multiclass microarray data. Given a microarray data with N tumor classes, n tumor samples and g genes per sample, one should identify a small subset of informative genes that contribute most to the prediction task. Various feature selection methods exist in literature. One way pointed in [1] is to use test statistics for the equality of the class means. Authors of [1] formulate first the expression levels of a given gene by a one-way analysis of variance model. Second, the power of genes in discriminating between tumor types is determined by a test statistic. The discrimination power is the value of the test evaluated at the expression level of the gene. The higher the discrimination power is, the more powerful the gene is in discriminating between tumor types. Thus, genes with higher power of discrimination are considered as informative genes.

Let Y_{jp} be the expression level from the pth sample of the jth class, the following general model is considered:

$$Y_{jp} = \mu_j + \epsilon_{jp} \quad \text{for } j = 1, \dots, N;\ p = 1, \dots, n_j \text{ with} \sum_{j=1}^{N} n_j = n. \tag{1}$$

In the model μ_j represents the mean expression level of the gene in class w_j, ϵ_{jp} are independent random variables and $E(\epsilon_{jp}) = 0$, $V(\epsilon_{jp}) = \sigma_j^2 < \infty$ for $j = 1, \dots, N$; $p = 1, \dots, n_j$.

For the case of homogeneity of variances, the ANOVA F or F test [16] is the optimal one testing the means equality hypothesis. With heterogeneity of variances, the task is challenging. However, it is known that, with a large number of genes present, usually in thousands, no practical test is available to locate the best set of genes. Thus, the authors of [1] studied six different statistics.

(i) ANOVA F test statistic, the definition of this test is

$$F = \frac{(n - N) \sum n_j \left(\overline{Y_{j\cdot}} - \overline{Y_{\cdot\cdot}}\right)^2}{(N - 1) \sum \left(n_j - 1\right) s_j^2}, \tag{2}$$

where $\overline{Y_{j\cdot}} = \sum_{p=1}^{n_j} Y_{jp}/n_j$ and $\overline{Y_{\cdot\cdot}} = \sum_{j=1}^{N} n_j \overline{Y_{j\cdot}}/n$, $s_j^2 = \sum_{p=1}^{n_j} (Y_{jp} - \overline{Y_{j\cdot}})^2/(n_j - 1)$. For simplicity, $\sum$ is used to indicate the sum taken over the index j. Under means equality hypothesis and assuming variance homogeneity, this test has a distribution of $F_{N-1,n-N}$ [16].

(ii) Brown-Forsythe test statistic [17], given by

$$B = \frac{\sum n_j \left(\overline{Y_{j\cdot}} - \overline{Y_{\cdot\cdot}}\right)^2}{\sum \left(1 - n_j/n\right) s_j^2}. \tag{3}$$

Under means equality hypothesis, B is distributed approximately as $F_{N-1,\tau}$ where

$$\tau = \frac{\left[\sum (1 - n_j/n) s_j^2\right]^2}{\sum \left(1 - n_j/n\right)^2 s_j^4 / \left(n_j - 1\right)}. \tag{4}$$

(iii) Welch test statistic [18], defined as

$$W = \frac{\sum \omega_j \left(\overline{Y_{j\cdot}} - \sum h_j \overline{Y_{j\cdot}}\right)^2}{(N-1) + 2(N-2)(N+1)^{-1} \sum \left(n_j - 1\right)^{-1} \left(1 - h_j\right)^2}, \tag{5}$$

with $\omega_j = n_j / s_j^2$ and $h_j = \omega_j / \sum \omega_j$. Under means equality hypothesis, W has an approximate distribution of F_{N-1,τ_ω} where

$$\tau_\omega = \frac{N^2 - 1}{3 \sum \left(n_j - 1\right)^{-1} \left(1 - h_j\right)^2}. \tag{6}$$

(iv) Adjusted Welch test statistic [19]. It is similar to Welch statistic and defined to be

$$W^* = \frac{\sum \omega_j^* \left(\overline{Y_{j\cdot}} - \sum h_j^* \overline{Y_{j\cdot}}\right)^2}{(N-1)+2(N-2)(N+1)^{-1} \sum \left(n_j-1\right)^{-1} \left(1 - h_j^*\right)^2}, \tag{7}$$

where $\omega_j^* = n_j / (\Phi_j s_j^2)$ with Φ_j chosen such that $1 \le \Phi_j \le (n_j - 1)/(n_j - 3)$ and $h_j^* = \omega_j^* / \sum \omega_j^*$. Under means equality hypothesis, W^* has an approximate distribution of F_{N-1,τ_ω^*} where

$$\tau_\omega^* = \frac{N^2 - 1}{3 \sum \left(n_j - 1\right)^{-1} \left(1 - h_j^*\right)^2}. \tag{8}$$

(v) Cochran test statistic [20]. This test statistic is simply the quantity appearing in the numerator of the Welch test statistic W, that is,

$$C = \sum \omega_j \left(\overline{Y_{j\cdot}} - \sum h_j \overline{Y_{j\cdot}}\right)^2. \tag{9}$$

Under means equality hypothesis, C has an approximate distribution of χ^2_{N-1}.

(vi) Kruskal-Wallis test statistic. This is the well-known nonparametric test given by

$$H = \frac{12}{n(n+1)} \sum \frac{R_j^2}{n_j} - 3(n+1), \tag{10}$$

where R_j is the rank sum for the jth class. The ranks assigned to Y_{jp} are those obtained from ranking the entire set of Y_{jp}. Assuming each $n_j \ge 5$, then under means equality hypothesis, H has an approximate distribution of χ^2_{N-1} [21].

These tests performances are evaluated and compared over different supervised learning methods applied to publicly available microarray datasets. Experimental results show that the model for gene expression values without assuming equal variances is more appropriate than that assuming equal variances. Besides, under heterogeneity of variances, Brown-Forsythe test statistic, Welch test statistic, adjusted Welch test statistic, and Cochran test statistic, perform much better than ANOVA F test statistic and Kruskal-Wallis test statistic.

2.2. Multitumor Classes with Selective Rejection. Once gene selection is processed, the classification problem should be solved. Let us define this diagnosis problem in the class-selective rejection scheme. Assuming that the multiclass cancer problem deals with N tumor classes noted $w_1 \dots w_N$ and that any patient or sample x belongs to one tumor class and has d informative genes, a decision rule consists in a partition Z of $\Re^d$ in I sets Z_i corresponding to the different decision options. In the simple classification scheme, the options are defined by the N tumor classes. In the class-selective rejection scheme, the options are defined by the N tumor classes and the subsets of tumor classes (i.e. assigning patient x to the subset of tumor classes $\{w_1, w_3\}$ means that x is assigned to cancer categories w_1 and w_3 with ambiguity).

The problem consists in finding the decision rule Z^* that minimizes a given loss function $c(Z)$ defined by

$$c(Z) = \sum_{i=1}^{I} \sum_{j=1}^{N} c_{ij} P_j P\left(D_i / w_j\right), \tag{11}$$

where c_{ij} is the cost of assigning a patient x to the ith decision option when it belongs to the tumor class w_j. The values of c_{ij} being relative since the aim is to minimize $c(Z)$, the values can be defined in the interval $[0;1]$ without loss of generality. P_j is the a priori probability of tumor class w_j and $P(D_i/w_j)$ is the probability that patients of the tumor class w_j are assigned to the ith option.

2.3. $\boldsymbol{\mu}$-1-SVM. To solve the multiclass diagnosis problem, an approach based on ν-1-SVM is proposed. Considering a set of m samples of a given tumor classes $X = \{x_1, x_2, \dots, x_m\}$ drawn from an input space $\mathcal{X}$, ν-1-SVM computes a decision function $f_X^\lambda(\cdot)$ and a real number b^λ in order to determine the region $\mathcal{R}^\lambda$ in $\mathcal{X}$ such that $f_X^\lambda(x) - b^\lambda \ge 0$ if the sample $x \in \mathcal{R}^\lambda$ and $f_X^\lambda(x) - b^\lambda < 0$ otherwise. The decision function $f_X^\lambda(\cdot)$ is parameterized by $\lambda = \nu m$ (with $0 \le \nu < 1$) to control the number of outliers. It is designed by minimizing the volume of $\mathcal{R}^\lambda$ under the constraint that all the samples of X, except the fraction ν of outliers, must lie in $\mathcal{R}^\lambda$. In order to determine $\mathcal{R}^\lambda$, the space of possible functions $f_X^\lambda(\cdot)$ is reduced to a Reproducing Kernel Hilbert Space (RKHS) with kernel function $K(\cdot,\cdot)$. Let $\Phi : \mathcal{X} \to \mathcal{H}$ be the mapping defined over the input space $\mathcal{X}$. Let $\langle \cdot,\cdot \rangle_{\mathcal{H}}$ be a dot product defined in $\mathcal{H}$. The kernel $K(\cdot,\cdot)$ over $\mathcal{X} \times \mathcal{X}$ is defined by:

$$\forall \left(x_p, x_q\right) \in \mathcal{X} \times \mathcal{X} \quad K\left(x_p, x_q\right) = \left\langle \Phi\left(x_p\right), \Phi\left(x_q\right) \right\rangle_{\mathcal{H}}. \tag{12}$$

Without loss of generality, $K(\cdot,\cdot)$ is supposed normalized such that for any $x \in \mathcal{X}$, $K(x,x) = 1$. Thus, all the mapped vectors $\Phi(x_p)$, $p = 1,\ldots,m$ are in a subset of a hypersphere with radius one and center O. Provided $K(\cdot,\cdot)$ is always positive and $\Phi(X)$ is a subset of the positive orthant of the hypersphere. A common choice of $K(\cdot,\cdot)$ is the Gaussian RBF kernel $K(x_p, x_q) = \exp[-1/2\sigma^2 \|x_p - x_q\|^2_{\mathcal{X}}]$ with σ the parameter of the Gaussian RBF kernel. ν-1-SVM consists of separating the mapped samples in $\mathcal{H}$ from the center O with a hyperplane $\mathcal{W}^\lambda$. Finding the hyperplane $\mathcal{W}^\lambda$ is equivalent to find the decision function $f_X^\lambda(\cdot)$ such that $f_X^\lambda(x) - b^\lambda = \langle w^\lambda, \Phi(x)\rangle_{\mathcal{H}} - b^\lambda \geq 0$ for the $(1-\nu)m$ mapped samples while $\mathcal{W}^\lambda$ is the hyperplane with maximum margin $b^\lambda/\|w^\lambda\|_{\mathcal{H}}$ with w^λ the normal vector of $\mathcal{W}^\lambda$.

This yields $f_X^\lambda(\cdot)$ as the solution of the following convex quadratic optimization problem:

$$\min_{w^\lambda, b^\lambda, \xi_p} \sum_{p=1}^{m} \xi_p - \lambda b^\lambda + \frac{\lambda}{2}\left\|w^\lambda\right\|^2_{\mathcal{H}}$$

$$\text{subject to} \left\langle w^\lambda, \Phi\left(x_p\right)\right\rangle_{\mathcal{H}} \geq b^\lambda - \xi_p, \quad \xi_p \geq \quad \forall p = 1,\ldots,m \tag{13}$$

where ξ_p are the slack variables. This optimization problem is solved by introducing lagrange multipliers α_p. As a consequence to Kuhn-Tücker conditions, w^λ is given by

$$w^\lambda = \frac{1}{\lambda}\sum_{p=1}^{m} \alpha_p \Phi\left(x_p\right), \tag{14}$$

which results in

$$f_X^\lambda(\cdot) - b^\lambda = \frac{1}{\lambda}\sum_{p=1}^{m} \alpha_p K\left(x_p, \cdot\right) - b^\lambda. \tag{15}$$

The dual formulation of (13) is obtained by introducing Lagrange multipliers as

$$\min_{\alpha_1,\ldots,\alpha_m} \frac{1}{2\lambda}\sum_{p=1}^{m}\sum_{q=1}^{m} \alpha_p^\lambda \alpha_q^\lambda K\left(x_p, x_q\right)$$

$$\text{with} \sum_{p=1}^{m} \alpha_p^\lambda = \lambda, \quad 0 \leq \alpha_p^\lambda \leq 1 \quad \forall p = 1,\ldots,m. \tag{16}$$

A geometrical interpretation of the solution in the RKHS is given by Figure 1. $f_X^\lambda(\cdot)$ and b^λ define a hyperplane $\mathcal{W}^\lambda$ orthogonal to $f_X^\lambda(\cdot)$. The hyperplane $\mathcal{W}^\lambda$ separates the $\Phi(x_p)$s from the sphere center, while having $b^\lambda/\|w^\lambda\|_{\mathcal{H}}$ maximum which is equivalent to minimize the portion $\mathcal{S}^\lambda$ of the hypersphere bounded by $\mathcal{W}^\lambda$ that contains the set $\{\Phi(x)$ s.t. $x \in \mathcal{R}^\lambda\}$.

Tuning ν or equivalently λ is a crucial point since it enables to control the margin error. It is obvious that changing λ leads to solve the optimization problem formulated in (16) in order to find the new region $\mathcal{R}^\lambda$. To obtain great computational savings and extend the search space of λ, we proposed to use ν-1-SVM regularization path [14, 15]. Regularization path was first introduced by Hastie et al. [14] for a binary SVM. Later, Rakotomamojy and Davy [15] developed the entire regularization path for a ν-1-SVM. The basic idea of the ν-1-SVM regularization path is that the parameter vector of a ν-1-SVM is a piecewise linear function of λ. Thus the principle of the method is to start with large λ, (i.e., $\lambda = m - \epsilon$) and decrease it towards zero, keeping track of breaks that occur as λ varies.

As λ decreases, $\|w^\lambda\|_{\mathcal{H}}$ increases and hence the distance between the sphere center and $\mathcal{W}^\lambda$ decreases. Samples move from being outside (non-margin SVs with $\alpha_p^\lambda = 1$ in Figure 1) to inside the portion $\mathcal{S}^\lambda$ (non-SVs with $\alpha_p^\lambda = 0$). By continuity, patients must linger on the hyperplane $\mathcal{W}^\lambda$ (margin SVs with $0 < \alpha_p^\lambda < 1$) while their α_p^λs decrease from 1 to 0. α_p^λs are piecewise-linear in λ. Break points occur when a point moves from a position to another one. Since α_p^λ is piecewise-linear in λ, $f^\lambda(\cdot)$ and b^λ are also piecewise-linear in λ. Thus, after initializing the regularization path (computing α_p^λ by solving (16) for $\lambda = m - \epsilon$), almost all the α_p^λs are computed by solving linear systems. Only for some few integer values of λ smaller than m, α_p^λs are computed by solving (16) according to [15].

Using simple linear interpolation, this algorithm enables to determine very rapidly the ν-1-SVM corresponding to any value of λ.

2.4. Multiclass SVM Based on μ-1-SVM. Given N classes and N trained ν-1-SVMs, one should design a supervised decision rule Z, moving from unary to multiclass classifier by assigning samples to a decision option. To determine the decision rule, first a prediction function should decide the winning option. A distance measure between x and the training class set w_j, using the ν-1-SVM parameterized by λ_j, is defined as follows:

$$d^{\lambda_j}(x) = \frac{\cos\left(\widehat{w^{\lambda_j}, \Phi(x)}\right)}{\cos\left(\theta^{\lambda_j}\right)} = \frac{\left\|w^{\lambda_j}\right\|_{\mathcal{H}}}{b^{\lambda_j}} \cos\left(\widehat{w^{\lambda_j}, \Phi(x)}\right), \tag{17}$$

where θ^{λ_j} is the angle delimited by w^{λ_j} and the support vector as shown in Figure 1. $\cos(\theta^{\lambda_j})$ is a normalizing factor which is used to make all the $d_j^\lambda(x)$ comparable.

Using $\|\Phi(x)\| = 1$ in (17) leads to the following:

$$d^{\lambda_j}(x) = \frac{\left\langle w^{\lambda_j}, \Phi(x)\right\rangle_{\mathcal{H}}}{b^{\lambda_j}} = \frac{1/\lambda_j \sum_{p=1}^{n_j} \alpha_p^{\lambda_j} K\left(x_p, x\right)}{b^{\lambda_j}}. \tag{18}$$

Since the $\alpha_p^{\lambda_j}$ are obtained by the regularization path for any value of λ_j, computing d^{λ_j} is considered as an easy-fast task. The distance measure $d^{\lambda_j}(x)$ is inspired from [22]. When data are distributed in a unimodal form, the $d^{\lambda_j}(x)$ is a decreasing function with respect to the distance between a sample x and the data mean. The probability density function is also a decreasing function with respect to the distance from the mean. Thus, $d^{\lambda_j}(x)$ preserves distribution order relations. In such case, and under optimality of the ν-1-SVM classifier, the use of $d^{\lambda_j}(x)$ should reach the same performances as the one obtained using the distribution.

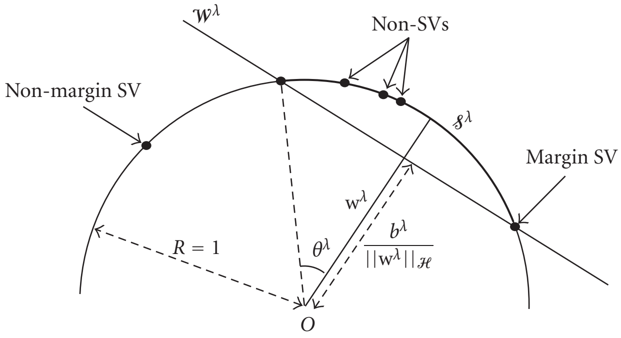


Figure 1: Training data mapped into the feature space on a portion $\mathcal{S}^\lambda$ of a hypersphere.

In the simplest case of multiclass problems where the loss function is defined as the error probability, a patient x is assigned to the tumor class maximizing $d^{\lambda_j}(x)$.

To extend the multiclass prediction process to the class-selective scheme, a weighted form of the distance measure is proposed. The weight β_j is associated to d^{λ_j}. β_j reflects an adjusted value of the distance d^{λ_j} according to the penalty associated with the tumor class w_j. Thus, introducing weights leads to treat differently each tumor class and helps solving problems with different costs c_{ij} on the classification decisions.

Finally, in the general case where the loss function is considered in the class-selective rejection scheme, the prediction process can be defined as follows: a blinded sample x is assigned to the ith option if and only if

$$\sum_{j=1}^{N} c_{ij} P_j \beta_j d^{\lambda_j}(x) \leq \sum_{j=1}^{N} c_{lj} P_j \beta_j d^{\lambda_j}(x), \quad \forall l = 1 \dots I,\ l \neq i. \tag{19}$$

Thus, in contrast to previous multiclass SVMs, which construct the maximum margin between classes and locate the decision hyperplane in the middle of the margin, the proposed approach resembles more to the robust Bayesian classifier. The distribution of each tumor class is considered and the optimal decision is slightly deviated toward the class with the smaller variance.

The proposed decision rule depends on $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\beta}$ vectors of σ_j, ν_j and β_j for $j = 1, \dots, N$. Tuning ν_j is the most time expensive task since changing ν_j leads to solve the optimization problem formulated in (16). Moreover, tuning ν_j is a crucial point, it enables to control the margin error. In fact, it was shown in [11] that this regularization parameter is an upper bound on the fraction of outliers and a lower bound on the fraction of the SVs. In [9, 23] a smooth grid search was supplied in order to choose the optimal values of $\boldsymbol{\nu}$. The N values ν_js were chosen equal to reduce the computational costs. However, this assumption reduces the search space of parameters too. To avoid this restriction, the proposed approach optimizes all the ν_j with $j = 1, \dots, N$ corresponding to the $N\nu$-1-SVMs using regularization path and consequently explores the entire parameters space. Thus the tuned ν_j are most likely to be the optimal ones. The parameter $\boldsymbol{\sigma}$ are set equals $\sigma_1 = \sigma_2 = \cdots = \sigma_N$.

The optimal vector of σ_j, λ_j and β_j, $j = 1, \dots, N$, is the one which minimizes an estimator of $c(Z)$ using a validation set. Since the problem is described by a sample set, an estimator $\hat{c}(Z)$ of $c(Z)$ given by (11) is used:

$$\hat{c}(Z) = \sum_{i=1}^{I} \sum_{j=1}^{N} c_{ij} \hat{P}_j \hat{P}\left(\frac{D_i}{w_j}\right), \tag{20}$$

where $\hat{P}_j$ and $\hat{P}(D_i/w_j)$ are the empirical estimators of P_j and $P(D_i/w_j)$, respectively.

The optimal rule is obtained by tuning λ_j, β_j and σ_j so that the estimated loss $\hat{c}(Z)$ computed on a validation set is minimum. This is accomplished by employing a global search for λ_j and β_j and an iterative search over the kernel parameter. For each given value σ of the parameter kernels, ν-1-SVMs are trained using the regularization path method on a training set. Then the minimization of $\hat{c}(Z)$ over a validation set is sought by solving an alternate optimization problem over λ_j and β_j which is easy since all ν-1-SVM solutions are easily interpolated from the regularization path. σ is chosen from a previously defined set of real numbers $[\sigma_0, \dots, \sigma_s]$ with $s \in \aleph$. Algorithm 1 elucidates the proposed approach.

3. Experimental Results

In this section, two experiments are reported in order to assess the performance of the proposed approach. First, the cancer diagnosis problem is considered in the traditional Bayesian framework. Five gene expression datasets and five supervised algorithms are considered. Each gene dataset was selected using the six test statistics of [1]. The decisions are given by the possible set of tumor classes and the loss function is defined as the probability of error to make results comparable with those of [1]. Second, in order to show the advantages of considering the multiclass cancer diagnosis in class-selective rejection scheme, one gene dataset is considered and studied with an asymmetric loss function. A cascade of classifiers with rejection options is used to ensure a reliable diagnosis. For both experiments, the loss function was minimized by determining the optimal parameters β_j and λ_j for $j = 1, \dots, N$ for a given kernel parameter $\boldsymbol{\sigma}$ and by testing different values of $\boldsymbol{\sigma}$ in the set $[2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2]$. Finally, the decision rule which minimizes the loss function is selected.

3.1. Bayesian Framework. Five multiclass gene expression datasets leukemia72 [24], ovarian [25], NCI [26, 27], lung cancer [28] and lymphoma [29] were considered. Table 1 describes the five genes datasets. For each dataset, the six test statistics F, B, W, W^*, C, and H were used to select informative genes.

The cancer diagnosis problem was considered in the traditional Bayesian framework. Decisions were given by the set of possible classes and loss function was defined by the error risk. This means that in (20) c_{ij} are defined according

```
$\boldsymbol{\theta} := \varnothing$
$C := \varnothing$
for $\sigma \leftarrow \sigma_0$ to $\sigma_s$ do
    /* **Using the Training Set** */
    for $j \leftarrow 1$ to $N$ do
        Train $\nu$-1-SVM on $w_j$, namely solving the QP (16)
        Derive the regularization path for $w_j$, namely compute the $\boldsymbol{\alpha}^{\lambda_j}$s
    end
    /* **Using the Validation Set** */
    $\boldsymbol{\lambda} := \boldsymbol{\lambda}_0$
    $\boldsymbol{\beta} := \boldsymbol{\beta}_0$
    repeat
        $d^{\lambda_j}(x) := \dfrac{1}{\lambda_j} \sum_{p=1}^{n_j} \alpha_p^{\lambda_j} K(x_p, x)/b^{\lambda_j}$
        $\widehat{P}_j := |w_j| / \sum_{j=1}^{N} |w_j|$        /* |  | = cardinality */
        Assign $x$ to a decision $\psi_i$ according to (19)
        $\widehat{P}(D_i/w_j) := |\{x \text{ of } w_j \text{ assigned to } \psi_i\}| / |\{x/x \in w_j\}|$
        $\widehat{c}(Z) := \sum_{i=1}^{I} \sum_{j=1}^{N} c_{ij} \widehat{P}_j \widehat{P}(D_i/w_j)$
        $\boldsymbol{\lambda} := \boldsymbol{\lambda}_{\text{new}}$/* construct the new vector according to the
        direction of greatest decrease */
        $\boldsymbol{\beta} := \boldsymbol{\beta}_{\text{new}}$
    until $\widehat{c}(Z)$ *is minimum*
    $\boldsymbol{\theta} := \boldsymbol{\theta} \cup \{\sigma, \boldsymbol{\lambda}, \boldsymbol{\beta}\}$
    $C := C \cup \{\widehat{c}(Z)\}$
end
index := min$\{C\}$
$\boldsymbol{\theta}_{\text{optimal}} := \boldsymbol{\theta}_{\text{index}}$
```

Algorithm 1: Multiclass SVM minimizing an asymmetric loss function.

Table 1: Multiclass gene expression datasets.

Dataset	Leukemia72	Ovarian	NCI	Lung cancer	Lymphoma
No. of gene	6817	7129	9703	918	4026
No. of sample	72	39	60	73	96
No. of class	3	3	9	7	9

Table 2: Loss function cost matrix in the Bayesian framework.

		Patient class				
		1	2	.	.	N
Prediction	1	0	1	·	·	1
	2	1	0	1		·
	·	·	·	·	·	·
	·	·		·	·	1
	N	1	·	·	1	0

to the Table 2. The performance of the proposed method was measured by evaluating its accuracy rate and it was compared to results obtained by the five predictors evoked in [1]: Naive Bayes, Nearest Neighbor, Linear Perceptron, Multilayer Perceptron Neural Network with five nodes in the middle layer, and Support Vector Machines with second-order polynomial kernel.

To compute the generalization accuracy of the proposed classifier, Leave One Out (LOO) resampling method is used to divide a gene dataset of n patients into two sets, a set of $n-1$ patients and a test set of 1 blinded patient. This method involves n separate runs. For each run, the first set of $n-1$ samples is divided using 5 Cross-validation (5-CV) into a training set and a validation set. $N\nu$-1-SVMs are trained using the training set for all values of ν_j. The decision is obtained by tuning the parameters β_j, λ_j and σ_j for $j = 1, \dots, N$ so that the loss function computed on the validation set is minimum. Optimal parameters are then used to build the decision rule using the whole $n-1$ samples. The blinded test set is classified according to this rule. The overall prediction error is the sum of the patients misclassified on all n runs.

Table 3 reports errors of the proposed algorithm, the average value and the median value of the 5 classifiers prediction errors reported in [1] when 50 informative genes are used. Table 4 reports values when 100 informative genes are used. F, B, W, $W*$, C, and H represent the six test statistics.

Experimental results show that, for ovarian, NCI, lung cancer and lymphoma multiclass genes problems, the proposed approach achieves competitive performances compared to the 5 classifiers reported in [1]. For these datasets, prediction errors of the proposed approach are less than the mean and median values of the 5 classifiers prediction errors reported in [1]. However, for leukemia72, the proposed

Table 3: Prediction errors of the proposed classifier, mean and median values of the **5** classifiers prediction errors according to [1] with 50 informative selected genes.

		F	*B*	*W*	*W**	*C*	*H*
Leukemia	Proposed algorithm	4	3	5	5	3	2
	Mean	3.4	2.4	2.8	2.8	3.2	3.0
	Median	3	2	3	3	3	3
Ovarian	Proposed algorithm	0	0	0	0	0	0
	Mean	0.2	0.0	0.0	0.0	0.0	0.0
	Median	0	0	0	0	0	0
NCI	Proposed algorithm	31	26	27	27	27	33
	Mean	36.0	32.0	27.4	26.0	27.0	35.4
	Median	35	29	27	27	27	35
Lung cancer	Proposed algorithm	14	16	16	16	16	15
	Mean	17.6	17.0	17.6	17.6	18.0	18.0
	Median	17	17	18	18	18	18
Lymphoma	Proposed algorithm	18	16	9	10	9	15
	Mean	23.8	19.8	14.0	14.0	12.8	22.0
	Median	23	19	12	12	13	20

Table 4: Prediction errors of the proposed classifier, mean and median values of the **5** classifiers prediction errors according to [1] with 100 informative selected genes.

		F	*B*	*W*	*W**	*C*	*H*
Leukemia	Proposed algorithm	5	2	3	3	4	6
	Mean	3.4	3.0	3.0	3.0	3.2	3.0
	Median	3	3	4	3	3	3
Ovarian	Proposed algorithm	0	0	0	0	0	0
	Mean	0.2	0.0	0.0	0.0	0.0	0.0
	Median	0	0	0	0	0	0
NCI	Proposed algorithm	33	21	26	25	26	36
	Mean	33.0	22.6	23.8	25.2	25.2	31.6
	Median	33	22	25	26	26	31
Lung cancer	Proposed algorithm	11	10	11	11	11	13
	Mean	12.2	12.2	11.4	12.2	12.2	15.8
	Median	12	12	11	11	11	14
Lymphoma	Proposed algorithm	16	16	11	10	11	17
	Mean	21.8	19.2	13.0	13.8	14.4	18.2
	Median	17	16	12	12	12	18

Table 5: Confusion matrix of 50 **W*** lung cancer dataset. Total of misclassified is equal to **16**.

		Patient class						
		Normal	SCLC	LCLC	SCC	AC2	AC3	AC1
Predicted decision	Normal	6	0	0	0	0	0	0
	SCLC	0	4	0	0	0	1	0
	LCLC	0	0	3	0	0	4	1
	SCC	0	0	0	16	0	3	0
	AC2	0	0	0	0	4	0	0
	AC3	0	1	1	0	1	4	0
	AC1	0	0	1	0	2	1	20

Table 6: Confusion Matrix of 50 *H* lung cancer dataset. Total of misclassified is equal to **15**.

		Patient class						
		Normal	SCLC	LCLC	SCC	AC2	AC3	AC1
Predicted decision	Normal	5	0	0	0	0	0	0
	SCLC	0	4	0	0	0	0	0
	LCLC	0	0	1	1	0	2	2
	SCC	0	0	2	14	0	1	0
	AC2	0	0	0	0	7	0	0
	AC3	0	0	2	1	0	8	0
	AC1	1	1	0	0	0	2	19

Table 7: Asymmetric cost matrix of the loss function.

		Patient class						
		Normal	SCLC	LCLC	SCC	AC2	AC3	AC1
Predicted decision	Normal	0	1	1	1	1	1	1
	SCLC	1	0	1	1	1	1	1
	LCLC	1	1	0	0.9	0.9	1	1
	SCC	1	1	0.9	0	0.9	1	0.9
	AC2	1	1	0.9	0.9	0	0.9	0.9
	AC3	1	1	0.9	0.9	0.9	0	0.9
	AC1	1	1	0.9	0.9	0.9	0.9	0
	{*LCLC, SCC, AC3*}	1	1	0.6	0.6	0.9	0.2	0.9
	All tumors	1	0.2	0.6	0.6	0.2	0.2	0.5
	All classes	0.6	0.2	0.6	0.6	0.2	0.6	0.6

Table 8: Confusion matrix of the 50 **W*** lung cancer problem with class-selective rejection using cost matrix defined in Table 7. Total of misclassified is equal to 10, total of partially and totally rejected samples is equal to 8.

		Patient class						
		Normal	SCLC	LCLC	SCC	AC2	AC3	AC1
Predicted decision	Normal	6	0	0	0	0	0	0
	SCLC	0	3	0	0	0	0	0
	LCLC	0	0	3	0	0	4	0
	SCC	0	0	0	16	0	2	0
	AC2	0	0	0	0	4	0	0
	AC3	0	0	0	0	1	3	0
	AC1	0	0	1	0	1	1	20
	{*LCLC, SCC, AC3*}	0	0	1	0	0	2	0
	All tumors	0	2	0	0	1	1	1
	All classes	0	0	0	0	0	0	0

algorithm performances are almost in the same range of those provided by the 5 classifiers reported in [1]. The proposed approach prediction error is equal, or in the worst case, slightly higher than the mean and median errors.

Moreover, we can note that focussing on the test statistics comparison, experimental results confirm those of [1]. B, W and W^* can be the most performing tests under variances heterogeneity assumptions.

3.2. Class-Selective Rejection Framework. In the following, we present the study of lung cancer problem in the class-selective rejection scheme. Lung cancer diagnosis problem is determined by the gene expression profiles of 67 lung tumors and 6 normal lung specimens from patients whose clinical course was followed for up to 5 years. The tumors comprised 41 Adenocarcinomas (ACs), 16 squamous cell carcinomas (SCCs); 5 cell lung cancers (LCLCs) and 5 small cell lung cancers (SCLCs). ACs are subdivided into three subgroups 21 AC of group 1 tumors, 7 AC of group 2 tumors and 13 AC of group 3 tumors. Thus, the multiclass diagnosis cancer consists of 7 classes.

Authors in [28] observed that AC of group 3 tumors shared strong expression of genes with LCLC and SCC tumors. Thus, poorly differentiated AC is difficult to distinguish from LCLC or SCC. Confusion matrices (Tables 5 and 6) computed in the Bayesian framework, with 50 W^* and 50 H prove well these claims. It can be noticed that 8 of the 16 misclassified 50 W^* patients and 8 of the 15 misclassified 50 H patients correspond to confusion between these three subcategories. Therefore, one may define a new decision option as a subset of these three classes to reduce error.

Moreover, same researches affirm that distinction between patients with nonsmall cell lung tumors (SCC, AC and LCLC) and those with small cell tumors or SCLC is extremely important, since they are treated very differently. Thus, a confusion or wrong decision among patients of nonsmall cell lung tumors should cost less than a confusion between nonsmall and small lung cells tumors. Besides, one may provide an extra decision option that includes all the subcategories of tumors to avoid this kind of confusion. Finally, another natural decision option can be the set of all

Table 9: Confusion matrix of the cascade classifier (50 $\mathbf{W}^*$ with rejection and 50 H classifier). Total of misclassified is equal to 13.

		Patient class						
		Normal	SCLC	LCLC	SCC	AC2	AC3	AC1
Predicted decision	Normal	6	0	0	0	0	0	0
	SCLC	0	4	0	0	0	0	0
	LCLC	0	0	3	0	0	4	1
	SCC	0	0	0	16	0	2	0
	AC2	0	0	0	0	5	0	0
	AC3	0	1	1	0	1	6	0
	AC1	0	0	1	0	1	1	20

classes, which means that the classifier has totally withhold taking a solution.

Given all these information, the loss function can be empirically defined according to the asymmetric cost matrix given in Table 7. Solving 50 W^* lung cancer problem in this scheme leads to the confusion matrix presented in Table 8. As a comparison with Table 5, one may mainly note that the number of misclassified patients decreases from 16 to 10 and 8 withhold decisions or rejected patients. This partial rejection contributes to avoid confusion between nonsmall and small lung cells tumors and reduces errors due to indistinctness among LCLC, SCC and AC3. Besides, according to the example under study, no patient is totally rejected. It is an expected result since initially (Table 5) there was no confusion between normal and tumor samples.

To take a decision concerning the rejected patients, we may refer to clinical analysis. It is worth to note that for partially rejected patients, clinical analysis will be less expensive in terms of time and money than those on completely blinded patients. Moreover, a supervised solution can be also proposed. It aims to use genes selected from another test statistic in order to assign rejected patients to one of the possible classes. According to Tables 3 and 4, prediction errors computed on same patients using genes selected by different test statistics may decrease since errors of two different test statistics do not occur on the same patients. Thus, we chose 50 H lung cancer dataset to reclassify the 8 rejected patients of Table 8. Five of them were correctly classified while three remained misclassified. Results are reported in Table 9. The number of misclassified patients decreases to 13 which is less than all the prediction errors obtained with 50 informative genes (lung cancer problem prediction errors of Table 3). In fact, many factors play an important role in the cascade classifiers system such as the asymmetric costs matrix which has been chosen empirically, the choice of test statistics, the number of classifiers in a cascade system,.... Such concerns are under study.

4. Conclusion

Cancer diagnosis using genes involve a gene selection task and a supervised classification procedure. This paper tackles the classification step. It considers the problem of gene-based multiclass cancer diagnosis in the general framework of class-selective rejection. It gives a general formulation of the problem and proposes a possible solution based on ν-1-SVM coupled with its regularization path. The proposed classifier minimizes any asymmetric loss function. Experimental results show that, in the particular case where decisions are given by the possible classes and the loss function is set equal to the error rate, the proposed algorithm, compared with the state of art multiclass algorithms, can be considered as a competitive one. In the class-selective rejection, the proposed classifier ensures higher reliability and reduces time and expense costs by introducing partial and total rejection. Furthermore, results prove that a cascade of classifiers with class-selective rejections can be considered as a good way to get improved supervised diagnosis. To get the most reliable diagnosis, the confusion matrix defining the loss function should be carefully chosen. Finding the optimal loss function according to performance constraints is an promising approach [30] which is actually under investigation.

5. Acknowledgments

The authors thank Dr. Dechang Chen of the Uniformed Services University of the Health Sciences for providing the gene selected data of leukemia72, ovarian, NCI, lung cancer and lymphoma using six test statistics ANOVA F, Brown-Forsythe test, Welch test, Adjusted Welch test, Cochran test and Kruskal-Wallis test. This work was supported by the "Conseil Régional Champagne Ardenne" and the "Fonds Social Européen".

References

[1] D. Chen, Z. Liu, X. Ma, and D. Hua, "Selecting genes by test statistics," *Journal of Biomedicine and Biotechnology*, vol. 2005, no. 2, pp. 132–138, 2005.

[2] S. Ramaswamy, P. Tamayo, R. Rifkin, et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 15149–15154, 2001.

[3] N. Jrad, E. Grail-Maës, and P. Beauseroy, "A supervised decision rule for multiclass problems minimizing a loss function," in *Proceedings of the 7th International Conference on Machine Learning and Applications (ICMLA '08)*, pp. 48–53, San Diego, Calif, USA, December 2008.

[4] T. M. Ha, "The optimum class-selective rejection rule," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 6, pp. 608–615, 1997.

[5] T. Horiuchi, "Class-selective rejection rule to minimize the maximum distance between selected classes," *Pattern Recognition*, vol. 31, no. 10, pp. 1579–1588, 1998.

[6] E. Grall-Maës, P. Beauseroy, and A. Bounsiar, "Multilabel classification rule with performance constraints," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, vol. 3, pp. 784–787, Toulouse, France, May 2006.

[7] N. Jrad, E. Grall-Maës, and P. Beauseroy, "Gaussian mixture models for multiclass problems with performance constraints," in *Proceedings of the 17th European Symposium on Artificial Neural Networks (ESANN '09)*, Bruges, Belgium, April 2009.

[8] L. Bottou, C. Cortes, J. Denker, et al., "Comparison of classifier methods: a case study in handwritten digit recognition," in *Proceedings of the 12th IAPR International Conference on Pattern Recognition (ICPR '94)*, vol. 2, pp. 77–82, Jerusalem, Israel, October 1994.

[9] X. Yang, J. Liu, M. Zhang, and K. Niu, "A new multiclass SVM algorithm based on one-class SVM," in *Proceedings of International Conference on Computational Science (ICCS '07)*, pp. 677–684, Beijing, China, May 2007.

[10] P.-Y. Hao and Y.-H. Lin, "A new multi-class support vector machine with multi-sphere in the feature space," in *Proceedings of the 20th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE '07)*, vol. 4570 of *Lecture Notes in Computer Science*, pp. 756–765, Kyoto, Japan, June 2007.

[11] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[12] D. Tax, *One-class classification: concept learning in the absence of counter-examples*, Ph.D. thesis, Technische Universiteit Delft, Delft, The Netherlands, 2001.

[13] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, Mass, USA, 2001.

[14] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for the support vector machine," *The Journal of Machine Learning Research*, vol. 5, pp. 1391–1415, 2004.

[15] A. Rakotomamojy and M. Davy, "One-class SVM regularization path and comparison with alpha seeding," in *Proceedings of the 15th European Symposium on Artificial Neural Networks (ESANN '07)*, pp. 271–276, Bruges, Belgium, April 2007.

[16] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, McGraw-Hill, New York, NY, USA, 5th edition, 2005.

[17] M. B. Brown and A. B. Forsythe, "The small sample behavior of some statistics which test the equality of several means," *Technometrics*, vol. 16, no. 1, pp. 129–132, 1974.

[18] B. L. Welch, "On the comparison of several mean values: an alternative approach," *Biometrika*, vol. 38, no. 3-4, pp. 330–336, 1951.

[19] J. Hartung, D. Argaç, and K. H. Makambi, "Small sample properties of tests on homogeneity in one-way Anova and meta-analysis," *Statistical Papers*, vol. 43, no. 2, pp. 197–235, 2002.

[20] W. G. Cochran, "Problems arising in the analysis of a series of similar experiments," *Journal of the Royal Statistical Society*, vol. 4, pp. 102–118, 1937.

[21] W. Daniel, *Biostatistics: A Foundation for Analysis in the Health Sciences*, John Wiley & Sons, New York, NY, USA, 1999.

[22] M. Davy, F. Desobry, A. Gretton, and C. Doncarli, "An online support vector machine for abnormal events detection," *Signal Processing*, vol. 86, no. 8, pp. 2009–2025, 2006.

[23] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[24] T. R. Golub, D. K. Slonim, P. Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[25] J. B. Welsh, P. P. Zarrinkar, L. M. Sapinoso, et al., "Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 3, pp. 1176–1181, 2001.

[26] D. T. Ross, U. Scherf, M. B. Eisen, et al., "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics*, vol. 24, no. 3, pp. 227–235, 2000.

[27] U. Scherf, D. T. Ross, M. Waltham, et al., "A gene expression database for the molecular pharmacology of cancer," *Nature Genetics*, vol. 24, no. 3, pp. 236–244, 2000.

[28] M. E. Garber, O. G. Troyanskaya, K. Schluens, et al., "Diversity of gene expression in adenocarcinoma of the lung," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13784–13789, 2001.

[29] A. A. Alizadeh, M. B. Elsen, R. E. Davis, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[30] E. Grall-Maës and P. Beauseroy, "Optimal decision rule with class-selective rejection and performance constraints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. 1, 2009.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 906865, 9 pages
doi:10.1155/2009/906865

Research Article

Combining Dissimilarities in a Hyper Reproducing Kernel Hilbert Space for Complex Human Cancer Prediction

Manuel Martín-Merino,[1] Ángela Blanco,[1] and Javier De Las Rivas[2]

[1] *Department of Computer Science, Universidad Pontificia de Salamanca (UPSA), C/Compañía 5, 37002 Salamanca, Spain*
[2] *Cancer Research Center (CIC-IBMCC, CSIC/USAL), Campus Miguel De Unamuno s/n, 37007 Salamanca, Spain*

Correspondence should be addressed to Manuel Martín-Merino, mmartinmac@upsa.es

Received 16 January 2009; Accepted 24 March 2009

Recommended by Dechang Chen

DNA microarrays provide rich profiles that are used in cancer prediction considering the gene expression levels across a collection of related samples. Support Vector Machines (SVM) have been applied to the classification of cancer samples with encouraging results. However, they rely on Euclidean distances that fail to reflect accurately the proximities among sample profiles. Then, non-Euclidean dissimilarities provide additional information that should be considered to reduce the misclassification errors. In this paper, we incorporate in the ν-SVM algorithm a linear combination of non-Euclidean dissimilarities. The weights of the combination are learnt in a (Hyper Reproducing Kernel Hilbert Space) HRKHS using a Semidefinite Programming algorithm. This approach allows us to incorporate a smoothing term that penalizes the complexity of the family of distances and avoids overfitting. The experimental results suggest that the method proposed helps to reduce the misclassification errors in several human cancer problems.

1. Introduction

DNA Microarray technology provides us a way to monitor the expression levels of thousands of genes simultaneously across a collection of related samples. This technology has been applied particularly to the prediction of different types of human cancer with encouraging results [1].

Support Vector Machines (SVM) [2] are powerful machine learning techniques that have been applied to the classification of cancer samples [3]. However, the categorization of different cancer types remains a difficult problem for classical SVM algorithms. In particular, the SVM is based on Euclidean distances that fail to reflect accurately the proximities among the sample profiles [4]. Non-Euclidean dissimilarities misclassify frequently different subsets of patterns because each one reflects complementary features of the data. Therefore, they should be integrated in order to reduce the fraction of patterns misclassified by the base dissimilarities.

In this paper, we introduce a framework to learn a linear combination of non-Euclidean dissimilarities that reflect better the proximities among the sample profiles. Each dissimilarity is embedded in a feature space using the Empirical Kernel Map [5, 6]. After that, learning the dissimilarity is equivalent to optimize the weights of the linear combination of kernels. Several approaches have been proposed to this aim. In [7, 8] the kernel is learnt optimizing an error function that maximizes the alignment between the input kernel and an idealized kernel. However, this error function is not related to the misclassification error and is prone to overfitting. To avoid this problem, [9] learns the kernel by optimizing an error function derived from the Statistical Learning Theory. This approach includes a term to penalize the complexity of the family of kernels considered. This algorithm is not able to incorporate infinite families of kernels and does not overcome the overfitting of the data.

In this paper, the combination of distances is learnt in a (Hyper Reproducing Kernel Hilbert Space) HRKHS following the approach of hyperkernels proposed in [10]. This formalism exhibits a strong theoretical foundation and is less sensitive to overfitting. Moreover, it allow us to work with infinite families of distances. The algorithm has been applied to the prediction of different kinds of human cancer. The experimental results suggest that the combination of dissimilarities in a Hyper Reproducing Kernel Hilbert Space improves the accuracy of classifiers based on a single distance, particularly for nonlinear problems. Besides, our approach outperforms the Lanckriet formalism specially for multicategory problems and is more robust to overfitting.

This paper is organized as follows. Section 2 introduces the algorithm proposed, the material and the methods employed. Section 3 illustrates the performance of the algorithm in the challenging problem of gene expression data analysis. Finally, Section 4 gets conclusions and outlines future research trends.

2. Material and Methods

2.1. Distances for Gene Expression Data Analysis. An important step in the design of a classifier is the choice of a proper dissimilarity that reflects the proximities among the objects. However, the choice of a good dissimilarity is not an easy task. Each measure reflects different features of the data and the classifiers induced by the dissimilarities misclassify frequently a different set of patterns. In this section, we comment shortly the main differences among several dissimilarities proposed to evaluate the proximity between biological samples considering their gene expression profiles. For a deeper description and definitions see [11].

Let $\mathbf{x} = [x_1, \dots, x_d]$ be the vectorial representation of a sample where x_i is the expression level of gene i. The *Euclidean distance* evaluates if the gene expression levels differ significantly across different samples:

$$d_{\text{euclid}}\,(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}. \tag{1}$$

An interesting alternative is the *cosine dissimilarity*. This measure will become small when the ratio between the gene expression levels is similar for the two samples considered. It differs significantly from the Euclidean distance when the data is not normalized by the L_2 norm:

$$d_{\text{cosine}}\,(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}. \tag{2}$$

The *correlation measure* evaluates if the expression level of genes change similarly in both samples. Correlation-based measures tend to group together samples whose expression levels are linearly related. The correlation differs significantly from the cosine if the means of the sample profiles are not zero. This measure is more sensitive to outliers:

$$d_{\text{cor}}\,(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{i=1}^{d} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{d} (x_i - \overline{x})^2}\sqrt{\sum_{j=1}^{d} (y_j - \overline{y})^2}}\ , \tag{3}$$

where $\overline{x}$ and $\overline{y}$ are the means of the gene expression profiles.

The *Spearman rank dissimilarity* is less sensitive to outliers because it computes a correlation between the ranks of the gene expression levels:

$$d_{\text{speam}}\,(x', y') = 1 - \frac{\sum_{i=1}^{d} (x'_i - \overline{x'})(y'_i - \overline{y'})}{\sqrt{\sum_{i=1}^{d} (x'_i - \overline{x'})^2}\sqrt{\sum_{i=1}^{d} (y'_i - \overline{y'})^2}}, \tag{4}$$

where $\mathbf{x}_{\mathbf{i}}^{'} = \text{rank}(\mathbf{x}_i)$ and $\mathbf{y}_j^{'} = \text{rank}(\mathbf{y}_j)$.

An alternative measure that helps to overcome the problem of outliers is the *Kendall-τ index* which is related to the Mutual Information probabilistic measure [11]:

$$d_{\text{kendall}}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{i=1}^{d}\sum_{j=1}^{d} C_{x_{ij}} - C_{y_{ij}}}{d(d-1)}\ , \tag{5}$$

where $C_{x_{ij}} = \text{sign}(x_i - x_j)$ and $C_{y_{ij}} = \text{sign}(y_i - y_j)$.

Finally, the dissimilarities have been transformed using the inverse multiquadratic kernel because this transformation helps to discover certain properties of the underlying structure of the data [12, 13]. The inverse multiquadratic transformation is based on the inverse multiquadratic kernel defined as follows:

$$k(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{\|\mathbf{x} - \mathbf{y}\|^2 + c^2}}, \tag{6}$$

where c is a smoothing parameter. Considering that $\|\mathbf{x} - \mathbf{y}\|$ is the Euclidean distance, (6) can be rewritten in terms of a dissimilarity as follows:

$$k(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{d_{ij}^2 + c^2}}. \tag{7}$$

The above nonlinear transformation gives more weight to small dissimilarities, particularly when c becomes small.

2.2. ν-Support Vector Machines. Support Vector Machines [2] are powerful classifiers that are able to deal with high dimensional and noisy data keeping a high generalization ability. They have been widely applied in cancer classification using gene expression profiles [1, 14]. In this paper, we will focus on the ν-Support Vector Machines (SVM). The ν-SVM is a reparametrization of the classical C-SVM [2] that allows to interpret the regularization parameter in terms of the number of support vectors and margin errors. This property helps to control the complexity of the approximating functions in an intuitive way. This feature is desirable for the application we are dealing with because the sample size is frequently small and the resulting classifiers are prone to overfitting.

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ be the training set codified in $\mathbb{R}^d$. We assume that each $\mathbf{x}_i$ belongs to one of the two classes labeled by $y_i \in \{-1, 1\}$. The SVM algorithm looks for the linear hyperplane $f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T\mathbf{x} + b$ that maximizes the margin $\gamma = 2/\|\mathbf{w}\|^2$. γ determines the generalization ability of the SVM. The slack variables ξ_i allow to consider classification errors and are defined as $\xi_i = \max\{0, 1 - y_i f(\mathbf{x}_i)\}$.

For the ν-SVM, the hyperplane that minimizes the prediction error is obtained solving the following optimization problem [2]:

$$
\begin{aligned}
&\min_{w,\{\xi_i\},\rho} \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{m}\sum_i \xi_i \\
&\text{s.t. } y_i(\langle \mathbf{w},\ \mathbf{x_i}\rangle + b) \geq \rho - \xi_i, \quad i = 1,\ldots,m, \\
&\xi_i \geq 0, \qquad \rho \geq 0 \quad i = 1,\ldots,m,
\end{aligned}
\tag{8}
$$

where ν is an upper bound on the fraction of margin errors and a lower bound on the number of support vectors. Therefore, this parameter controls the complexity of the approximating functions.

The optimization problem can be solved efficiently in the dual space and the discriminant function can be expressed exclusively in terms of scalar products:

$$
f(\mathbf{x}) = \sum_{\alpha_i > 0} \alpha_i y_i \langle \mathbf{x}, \mathbf{x_i}\rangle + b, \tag{9}
$$

where α_i are the Lagrange multipliers in the dual optimization problem. The ν-SVM algorithm can be easily extended to the nonlinear case substituting the scalar products by a Mercer kernel [2]. Besides, non-Euclidean dissimilarities can be incorporated into the ν-SVM via the kernel of dissimilarities [5].

Finally, several approaches have been proposed in the literature to extend the SVM to deal with multiple classes. In this paper, we have followed the one-against-one (OVO) strategy. Let k be the number of classes, in this approach $k(k-1)/2$ binary classifiers are trained and the appropriate class is found by a voting scheme. This strategy compares favorably with more sophisticated methods and it is more efficient computationally than the one-against-rest (OVR) approach [15].

2.3. Empirical Kernel Map. The Empirical Kernel Map allows us to incorporate non-Euclidean dissimilarities into the SVM algorithm using the kernel trick [5, 13].

Let $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a dissimilarity and $R = \{p_1,\ldots,p_n\}$ a subset of representatives drawn from the training set. Define the mapping $\phi : \mathcal{F} \to \mathbb{R}^n$ as

$$
\phi(z) = D(z,R) = [d(z,p_1), d(z,p_2),\ldots,d(z,p_n)]. \tag{10}
$$

This mapping defines a dissimilarity space where feature i is given by $d(\cdot,p_i)$.

The set of representatives R determines the dimensionality of the feature space. The choice of R is equivalent to select a subset of features in the dissimilarity space. Due to the small number of samples in our application, we have considered the whole training set as representatives. Notice that it has been suggested in literature [13] that for small samples reducing the set of representatives does not help to improve the classifier performance.

2.4. Learning a Linear Combination of Dissimilarities in an HRKHS. In order to learn a linear combination of non-Euclidean dissimilarities, we follow the approach of Hyperkernels developed by [10]. To this aim, each distance is embedded in an RKHS via the Empirical Kernel Map presented in Section 2.3. Next, a regularized quality functional is introduced that incorporates an l_2-penalty over the complexity of the family of distances considered. The solution to this regularized quality functional is searched in a Hyper Reproducing Kernel Hilbert Space. This allows to minimize the quality functional using an SDP approach.

Let $X_{\text{train}} = \{x_1, x_2,\ldots,x_m\}$ and $Y_{\text{train}} = \{y_1, y_2,\ldots,y_n\}$ be a finite sample of training patterns where $y_i \in \{-1,+1\}$. Let $\mathcal{K}$ be a family of semidefinite positive kernels. Our goal is to learn a kernel of dissimilarities $k \in \mathcal{K}$ that represents the combination of dissimilarities and minimizes the following empirical quality functional:

$$
Q_{\text{emp}}(f, X_{\text{train}}, Y_{\text{train}}) = \frac{1}{m}\sum_{i=1}^{m} l(x_i, y_i, f(x_i)) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}}, \tag{11}
$$

where l is a loss function, $\|\ \|_{\mathcal{H}}$ is the L_2 norm defined in a reproducing kernel Hilbert space, and λ is a regularization parameter that controls the balance between training error and the generalization ability.

By virtue of the representer theorem [2], we know that (11) can be written as a kernel expansion:

$$
Q_{\text{emp}} = \min_{\alpha, k}\left[\frac{1}{m}\sum_{i=1}^{m} l(x_i, y_i, [K\alpha]_i) + \frac{\lambda}{2}\alpha^T K\alpha\right]. \tag{12}
$$

However, if the family of kernels $\mathcal{K}$ is complex enough it is possible to find a kernel that achieves zero error overfitting the data. To avoid this problem, we introduce a term that penalizes the kernel complexity in an HRKHS. A rigorous definition of the HRKHS is provided in the appendix:

$$
Q_{\text{reg}}(k, X, Y) = Q_{\text{emp}}(k, X, Y) + \frac{\lambda_Q}{2}\|k\|^2_{\underline{H}}, \tag{13}
$$

where $\|\ \|_{\mathcal{H}}$ is the L_2 norm defined in the Hyper Reproducing Kernel Hilbert space generated by the hyperkernel $\underline{k}$. λ_Q is a regularization parameter that controls the complexity of the resulting kernel.

The following theorem allows us to write the solution to the minimization of this regularized quality functional as a linear combination of hyperkernels in an HRKHS.

Theorem 1 (Representer theorem for Hyper-RKHS [10]). *Let X, Y be the combined training and test set, then each minimizer $k \in \underline{\mathcal{H}}$ of the regularized quality functional $Q_{reg}(k, X, Y)$ admits a representation of the form*

$$
k(x, x') = \sum_{i,j=1}^{m} \beta_{ij}\underline{k}\left(\left(x_i, x_j\right), (x, x')\right), \tag{14}
$$

for all $x,\ x' \in X$, where $\beta_{ij} \in \mathbb{R}$, for each $1 \leq i, j \leq m$.

However, we are only interested in solutions that give rise to positive semidefinite kernels. The following condition over the hyperkernels [10] allows us to guarantee that the solution is a positive semidefinite kernel.

Property 1. *Given a hyperkernel $\underline{k}$ with elements such that for any fixed $\underline{x} \in \underline{X}$, the function $k(x_p, x_q) = \underline{k}(\underline{x}, (x_p, x_q))$, with $x_p, x_q \in \mathcal{X}$, is a positive semidefinite kernel, and $\beta_{ij} \geq 0$ for all $i, j = 1, \dots, m$, then the kernel*

$$k\left(x_p, x_q\right) = \sum_{i,j=1}^{m} \beta_{ij} \underline{k}\left(x_i, x_j, x_p, x_q\right) \tag{15}$$

is positive semidefinite.

Now, we address the problem of combining a finite set of dissimilarities. As we mentioned in Section 2.3, each dissimilarity can be represented by a kernel using the Empirical Kernel Map. Next, the hyperkernel is defined as

$$\underline{k}(\underline{x}, \underline{x'}) = \sum_{i=1}^{n} c_i k_i(\underline{x}) k_i(\underline{x'}), \tag{16}$$

where each k_i is a positive semidefinite kernel of dissimilarities and c_i is a constant ≥ 0.

Now, we show that $\underline{k}$ is a valid hyperkernel. First, $\underline{k}$ is a kernel because it can be written as a dot product $\langle \underline{\Phi}(\underline{x}), \underline{\Phi}(\underline{x'}) \rangle$ where

$$\underline{\Phi}(\underline{x}) = (\sqrt{c_1}\ k_1(\underline{x}), \sqrt{c_2}\ k_2(\underline{x}), \dots, \sqrt{c_n}\ k_n(\underline{x})). \tag{17}$$

Next, the resulting kernel (15) is positive semidefinite because for all $\underline{x}$, $\underline{k}(\underline{x}, (x_p, x_q))$ is a positive semidefinite kernel and β_{ij} can be constrained to be ≥ 0. Besides, the linear combination of kernels is a kernel and therefore is positive semidefinite. Notice that $\underline{k}(\underline{x}, (x_p, x_q))$ is positive semidefinite if $c_i \geq 0$ and k_i are pointwise positive for training data. Both RBF and multiquadratic kernels verify this condition.

Finally, we show that the resulting kernel is a linear combination of the original k_i. Substituting the expression of the hyperkernel (16) in (15), the kernel is written as

$$k\left(x_p, x_q\right) = \sum_{i,j=1}^{m} \beta_{ij} \sum_{l=1}^{n} c_l k_l\left(x_i, x_j\right) k_l\left(x_p, x_q\right). \tag{18}$$

Now the kernel can be written as a linear combination of base kernels:

$$k\left(x_p, x_q\right) = \sum_{l=1}^{n} \left[c_l \sum_{i,j=1}^{m} \beta_{ij} k_l\left(x_i, x_j\right) \right] k_l\left(x_p, x_q\right). \tag{19}$$

Therefore, the above kernel introduces into the ν-SVM a linear combination of base dissimilarities represented by k_l with coefficients $\gamma_l = c_l \sum_{i,j=1}^{m} \beta_{ij} k_l(x_i, x_j)$.

The previous approach can be extended to an infinite family of distances. In this case, the space that generates the kernel is infinite dimensional. Therefore, in order to work in this space, it is necessary to define a hyperkernel and to optimize it using an HRKHS. Let k be a kernel of dissimilarities. The hyperkernel is defined as follows [10]:

$$\underline{k}(\underline{x}, \underline{x'}) = \sum_{i=0}^{\infty} c_i (k(\underline{x}) k(\underline{x'}))^i, \tag{20}$$

where $c_i \geq 0$ and $i = 0, \dots, \infty$. In this case, the nonlinear transformation to feature space is infinite dimensional. Particularly, we are considering all powers of the original kernels which is equivalent to transform nonlinearly the original dissimilarities:

$$\underline{\Phi}(\underline{x}) = \left(\sqrt{(c_1)} k(\underline{x}), \sqrt{(c_2)} k^2(\underline{x}), \dots, \sqrt{(c_n)} k^n(\underline{x}) \right), \tag{21}$$

where n is the dimensionality of the space which is infinite in this case. As we mentioned in Section 2.1, nonlinear transformations of a given dissimilarity provide additional information that may help to improve the classifier performance.

As for the finite family, it can be easily shown that $\underline{k}$ is a valid hyperkernel provided that the kernels considered are pointwise positive. The Inverse Multiquadratic kernel satisfies this condition. Next, we derive the hyperkernel expression for the multiquadratic kernel.

Proposition 1 (see [Harmonic Hyperkernel]). *Suppose k is a kernel with range* $[0, 1]$ *and* $c_i = (1 - \lambda_h)\lambda_h^i$, $i \in \mathbb{N}$, $0 < \lambda_h < 1$. *Then, computing the infinite sum in* (20), *one has the following expression for the harmonic hyperkernel:*

$$\underline{k}(\underline{x}, \underline{x'}) = (1 - \lambda_h) \sum_{i=0}^{\infty} (\lambda_h k(\underline{x}) k(\underline{x'}))^i = \frac{1 - \lambda_h}{1 - \lambda_h k(\underline{x}) k(\underline{x'})}, \tag{22}$$

λ_h is a regularization term that controls the complexity of the resulting kernel. Particularly, larger values for λ_h give more weight to strongly nonlinear kernels while smaller values give coverage for wider kernels.

In this paper one has considered the inverse multiquadratic kernel defined in (6). *Substituting in* (22), *one gets the inverse multiquadratic hyperkernel:*

$$\underline{k}(\underline{x}, \underline{x\prime}) = \frac{1 - \lambda_h}{1 - \lambda_h \left(\left(\|x - x'\|^2 + c^2 \right) \left(\|x'' - x'''\|^2 + c^2 \right) \right)^{-1/2}}, \tag{23}$$

where $\underline{x} = (x, x')$ and $\underline{x'} = (x'', x''')$.

2.5. ν-SVM in an HRKHS. In this section, we detail how to learn the kernel for a ν-Support Vector Machine in an HRKHS. First, we will introduce the optimization problem and next, we will explain shortly how to solve it using an SDP approach.

We start some notation that is used in the ν-SVM algorithm. For $p, q, r \in \mathbb{R}^n$, $n \in \mathbb{N}$ let $r = p \circ q$ be defined as element by element multiplication, $r_i = p_i \times q_i$. The pseudoinverse of a matrix K is denoted by $K^{\dagger}$. Define the hyperkernel Gram matrix $\underline{K}$ by $\underline{K}_{ijpq} = \underline{k}((x_i, x_j), (x_p, x_q))$, the kernel matrix $K = \text{reshape}(\underline{K}\beta)$ (reshaping an m^2 by 1 vector, $\underline{K}\beta$, to an $m \times m$ matrix), $Y = \text{diag}(y)$ (a matrix with y on the diagonal and zero otherwise), $G(\beta) = YKY$ (the dependence on β is made explicit), and 1 is a vector of ones.

The ν-SVM considered in this paper uses an l_1 soft margin, where $l(x_i, y_i, f(x_i)) = \max(0, 1 - y_i f(x_i))$. This error is less sensitive to outliers which are convenient features for microarray datasets. Let ξ_i be the slack variables that allow for errors in the training set. Substituting in (13) Q_{emp} by the one optimized by ν-SVM (8) the regularized quality functional in an HRKHS can be written as

$$\begin{aligned}&\min_{k\in\underline{H}}\min_{\mathbf{w}\in\mathcal{H}_k}\ \frac{1}{m}\sum_{i=1}^{m}\xi_i+\frac{1}{2}\|\mathbf{w}\|^2_{\mathcal{H}}-\nu\rho+\frac{\lambda_Q}{2}\|k\|^2_{\underline{H}}\\ &\text{s.t.}\quad y_i f(x_i)\geq\rho-\xi_i,\quad i=1,\ldots,m,\\ &\qquad\ \xi_i\geq 0\quad i=1,\ldots,m,\end{aligned}\tag{24}$$

where ν is the regularization parameter that achieves a balance between training error and the complexity of the approximating functions and λ_Q is a parameter that penalizes the complexity of the family of kernels considered. The minimization of the previous equation leads to the following SDP optimization problem [10].

$$\min_{\beta,\gamma,\eta,\xi,\chi}\ \frac{1}{2}t_1-\chi\nu+\frac{1}{m}\xi^T1+\frac{\lambda_Q}{2}t_2\tag{25}$$

$$\text{s.t.}\quad \chi\geq 0,\ \eta\geq 0,\ \xi\geq 0,\ \beta\geq 0,\tag{26}$$

$$\left\|\underline{K}^{1/2}\beta\right\|\leq t_2,\ 1^T\beta=1,\tag{27}$$

$$\begin{bmatrix}G(\beta) & z\\ z^T & t_1\end{bmatrix}\succeq 0,\tag{28}$$

where $z=\gamma y+\chi 1+\eta-\xi$

The value of α which optimizes the corresponding Lagrange function is $G(\beta)^{\dagger}z$, and the classification function, $f=\text{sign}(K(\alpha\circ y)-b_{\text{offset}})$, is given by

$$f=\text{sign}\left(KG(\beta)^{\dagger}(y\circ z)-\gamma\right),\tag{29}$$

$\underline{K}$ is the hyperkernel defined in Section 2.4 which represents the combination of dissimilarities considered. Finally, the algorithm proposed can be easily extended to deal with multiple classes via a one-against-one approach (OVO). This strategy is simple, more efficient computationally than the OVR, and compares well with more sophisticated multicategory SVM methods [15].

2.6. Implementation. The optimization problem (25) were solved using SeDuMi 1.1R3 [16] and YALMIP [17] SDP optimization packages running under MATLAB.

As in the SDP problem there are m^2 coefficients β_{ij}, the computational complexity is high. However, it can be significantly reduced if the Hyperkernel $\{\underline{k}((x_i,x_j),\cdot)\mid 1\leq i,j\leq m^2\}$ is approximated by a small fraction of terms, $p\ll m^2$ for a given error. In particular, we have chosen an $m\times p$ truncated lower triangular matrix G which approximate the hyperkernel matrix to an error $\delta=10^{-6}$ using the incomplete Cholesky factorization method [18].

2.7. Datasets and Preprocessing. The gene expression datasets considered in this paper correspond to several human cancer problems and exhibit different features as shown in Table 1. We have considered both, binary and multicategory problems with a broad range of signal to noise ratio (Var/Samp.), different number of samples, and varying priors for the larger category. All the datasets are available from the Broad Institute of MIT and Harvard http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi/. Next we detail the features and preprocessing applied to each dataset.

Table 1: Features of the different cancer datasets

	Clases	Samples	Genes	Var/Samp.	Priors %
Lymphoma DLBCL	2	77	6817	88	75.3
Lymphoma MLBCL/DLBCL	2	210	44928	213	84
Breast cancer LN	2	49	7129	145	51
Medulloblastoma	2	60	7129	119	65
Breast cancer B	3	49	1213	24.7	52
DLBCL survival C	4	58	3795	65.4	27
DLBCL survival D	4	129	3795	29.4	38

The first dataset was obtained from 77 patients with (diffuse large B-cell lymphoma) DLBCL (58 samples) or FL (follicular lymphoma) (19 samples) and they were subjected to transcriptional profiling using oligonucleotide Affymetrix gene chip *hu*68000 containing probes for 6817 genes [19]. The second dataset consists of frozen tumors specimens from newly diagnosed, previously untreated MLBCL patients (34 samples) and DLBCL patients (176 samples). They were hybridized to Affymetrix *hgu*133*b* gene chip containing probes for 44000 genes [20]. In both cases the raw intensities have been normalized using the rma algorithm [21] available from Bioconductor package [11]. The third problem we address concerns the clinically important issue of metastatic spread of the tumor. The determination of the extent of lymph node involvement in primary breast cancer is the single most important risk factor in disease outcome and here the analysis compares primary cancers that have not spread beyond the breast to ones that have metastasized to axillary lymph nodes at the time of diagnosis. We identified tumors as “reported negative” (24) when no positive lymph nodes were discovered and “reported positive” (25) for tumors with at least three identifiably positive nodes [22]. All assays used the human HuGeneFL Genechip microarray containing probes for 7129 genes. The fourth dataset [23] address the clinical challenge concerning medulloblastoma due to the variable response of patients to therapy. Whereas some patients are cured by chemotherapy and radiation, others have progressive disease. The dataset consists of 60 samples containing 39 medulloblastoma survivors and 21 treatment failures. Samples were hybridized to Affymetrix HuGeneFL arrays containing 5920 known genes and 897 expressed sequence tags.

All the datasets have been standarized subtracting the median and dividing by the Inter-quantile range. The rescaling were performed based only on the training set to avoid bias.

Regarding the identification of multiple classes of cancer we have considered three different datasets. The first one consists of 49 samples of Breast Cancer generated using 1-channel oligonucleotide Affymetrix HuGeneFl [1]. The second and third datasets consist of 58 and a129 samples from Diffuse large B-cell lymphoma with survival data. Fourth different subclasses can be identified. Data preparatory steps have been performed by the authors of the primary study [1]. The 10% oligonucleotides with smaller Interquantile Range were filtered to remove genes with expression level constant across samples.

2.8. Performance Evaluation. In order to assure an honest evaluation of all the classifiers we have performed a double loop of crossvalidation [15]. The outer loop is based on stratified tenfold cross-validation that iteratively splits the data in ten sets, one for testing and the others for training. The inner loop perform stratified ninefold cross-validation over the training set and is used to estimate the optimal parameters avoiding overfitting. The stratified variant of cross-validation keeps the same proportion of patterns for each class in training and test sets. This is necessary in our problem because the class proportions are not equal. Finally, the error measure considered to evaluate the classifiers has been accuracy. This metric computes the proportion of samples misclassified. The accuracy is easy to interpret and allows us to compare with the results obtained by previously published studies.

2.9. Parameters for the Classification Algorithm. The parameters for the ν-SVM and for the classifiers based on a linear combination of dissimilarities have been set up by a nested stratified tenfold crossvalidation procedure [15]. This method avoids the overfitting as is described in Section 2.8 and takes into account the asymmetric distribution of class priors.

For the ν-SVM we have considered both, linear and inverse multiquadratic kernels. The optimal parameters have been obtained by a grid search strategy over the following set of values: $\nu = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\sigma = \{d/2, d, 2d\}$, where d denotes the dimensionality of the input space.

Additionally, for the finite family of distances $c_i = 1/M$ where M is the number of dissimilarities considered, and $\lambda_Q = 1$ because the misclassification errors are hardly sensitive to the regularization parameter that controls the kernel complexity. Finally, for the infinite family of dissimilarities, the regularization parameter λ_h in the Harmonic hyperkernel (22) has been set up to 0.6 which gives an adequate coverage of various kernel widths. Smaller values emphasizes only wide kernels. All the base kernel of dissimilarities have been normalized so that all ones have the same scale.

Regarding the Lanckriet [9] formalism that allows to combine a finite set of dissimilarities, several values for the regularization parameter C have been tried, $C = \{0.1, 1, 10, 100, 1000\}$. A grid search strategy has been applied to determine the best values for both, the kernel parameters and the regularization parameter. The kernel matrices have been normalized by the trace as recommended in the original paper.

Table 2: Accuracy for the ν-SVM using a linear combination of non-Euclidean dissimilarities in an HRKHS. The ν-SVM based on the best distance and coordinates and the Lanckriet formalism have been taken as a reference.

Technique	Limphoma	Limphoma cell B	Breast LN	Brain
ν-SVM (coordinates)	6.66%	7.14%	8.16%	16.6%
ν-SVM (best distance)	6.66%	5.71%	8.16%	13.3%
ν-SVM (nonlinear kernel)	6.25%	5.71%	8.16%	11.6%
Lanckriet (finite family)	5%	7.62%	8.16%	11.67%
Finite family of distances	5%	7.14%	10%	10%
Infinite family of distances	5%	5.71%	8%	8.33%

2.10. Gene Selection. Gene selection can improve significantly the classifier performance [24]. Therefore, we have evaluated the classifiers for the following subsets of genes $\{280, 146, 101, 56, 34\}$. The ν-SVM is robust against noise and is able to deal with high dimensional data. However, the empirical evidence suggests that considering a larger subset of genes or even the whole set of genes increases the misclassification errors.

The genes are ranked according to the ratio of between-group to within-group sums of squares defined in [25]:

$$BW(j) = \frac{\sum_i \sum_k I(y_i = k)\left(\overline{x}_{\cdot j}^{(k)} - \overline{x}_{\cdot j}\right)^2}{\sum_i \sum_k I(y_i = k)\left(x_{ij} - \overline{x}_{\cdot j}^{(k)}\right)^2}, \qquad (30)$$

where $\overline{x}_{\cdot j}^{(k)}$ and $\overline{x}_{\cdot j}$ denote "respectively" the average expression level of gene j for class k and the overall average expression level of gene j across all samples, y_i denotes the class of sample i, and $I(\cdot)$ is the indicator function. Next, the top ranked genes are chosen. This feature selection method is simple but compares well with more sophisticated methods [24]. Finally, the ranking of genes has been carried out considering only the training set to avoid bias. Therefore, feature selection is repeated in each iteration of cross-validation.

3. Results and Analysis

The algorithms proposed have been applied to the identification of several cancer human samples using microarray gene expression data.

First, we address several binary categorization problems.

Table 2 reports the accuracy for the two combination approaches proposed in this paper. The first one considers the finite set of dissimilarities introduced in Section 2.1. The second one considers an infinite family of distances obtained by transforming nonlinearly the base dissimilarities

Table 3: Accuracy for the ν-SVM using a linear combination of non-Euclidean dissimilarities in an HRKHS. The ν-SVM based on the best distance, the classical ν-SVM, and the Lanckriet formalism have been taken as a reference.

Technique	Breast B	DLBCL C	DLBCL D
ν-SVM (Coordinates)	10.20%	6.89%	12.96%
ν-SVM (Best Distance)	8.6%	6.89%	14.81%
ν-SVM (Nonlinear kernel)	8.16%	6.89%	12.96%
Lanckriet (finite family)	8%	10.3%	25.2%
Infinite family of distances	6%	5.33%	16%

to feature space. We have compared with the ν-SVM based on the best distance (linear and nonlinear kernel) and the classical ν-SVM. The performance for the Lanckriet formalism [9] that allow us to incorporate a finite linear combination of dissimilarities is also reported.

Before computing the kernel of dissimilarities, all the distances have been transformed using the multiquadratic kernel introduced in Section 2.1. This nonlinear transformation helps to improve the accuracy for all the techniques evaluated. From the analysis of Table 2, the following conclusions can be drawn.

(i) The ν-SVM based on a finite set of distances improves the ν-SVM based on the best dissimilarity for brain prognosis and Lymphoma datasets. The error is not reduced for Lymphoma cell B and Breast LN. This may be explained because the ratio (var/samp.) in Table 1 suggests that both datasets are quite noisy and nonlinear. The combination of a finite set of dissimilarities is not able to improve the separation between classes and increases slightly the overfitting of the data. Similarly, our algorithm helps to improve the SVM based on coordinates, particularly for the previous problems. We also report that working directly from a dissimilarity matrix may help to reduce the misclassification errors.

(ii) The infinite family of distances outperforms the ν-SVM based on the best distance disregarding the kernel considered for all the datasets. The improvement is more relevant in brain cancer prognosis. Brain cancer prognosis is a complex problem according to the original study [23] and the nonlinear transformations of the dissimilarities help to reduce the misclassification errors. Besides, the infinite family improves the accuracy of the finite family of distances particularly for lymphoma cell B and Breast LN. This suggests that both datasets are nonlinear.

(iii) The Lanckriet formalism and the finite family of dissimilarities perform similarly. However, the infinite family of distances outperforms the Lanckriet formalism particularly for brain and Lymphoma cell B which are more complex problems.

(iv) The best distance depends on the dataset considered.

Next we move to the categorization of multiple cancer types.

Table 3 compares the proposed algorithms with ν-SVM based on the best distance (linear and nonlinear kernel) and the classical ν-SVM. The accuracy for the Lanckriet formalism has also been reported. Our approach considers an infinite family of distances obtained by transforming nonlinearly the base dissimilarities to feature space.

Before computing the kernel of dissimilarities, all the distances have been transformed using the multiquadratic kernel introduced in Section 2.1. From the analysis of Table 3, the following conclusions can be drawn.

(i) The combination of non-Euclidean dissimilarities helps to improve the SVM based on the best dissimilarity disregarding the kernel considered for the two first datasets. The error is slightly larger for the third dataset which may suggest that the problem is linear.

(ii) Our algorithm improves the SVM based on coordinates. The experimental results suggest that the nonlinear transformations of the dissimilarities help to increase the separation among classes.

(iii) The Hyperkernel classifier outperforms the Lanckriet formalism for multicategory problems. As the number of classes growths the number of samples per class comes down and the Lanckriet formalism seems to be less robust to overfitting.

Finally, notice that our algorithm allow us to work with applications in with only a dissimilarity is defined. Moreover, we avoid the complex task of choosing a dissimilarity that reflects properly the proximities among the sample profiles.

4. Conclusions

In this paper, we propose two methods to incorporate in the ν-SVM algorithm a linear combination of non-Euclidean dissimilarities. The family of distances is learnt in a (Hyper Reproducing Kernel Hilbert Space) HRKHS using a Semidefinite Programming approach. A penalty term has been added to avoid the overfitting of the data. The algorithm has been applied to the classification of complex cancer human samples. The experimental results suggest that the combination of dissimilarities in a Hyper Reproducing Kernel Hilbert Space improves the accuracy of classifiers based on a single distance particularly for nonlinear problems. Besides, this approach outperforms the Lanckriet formalism specially for multi-category problems and is more robust to overfitting. Future research trends will focus on learning the combination of dissimilarities for other classifiers such as k-NN.

Appendix

In this section we define rigorously the Hyper-Reproducing Kernel Hilbert Spaces. First, we define a Reproducing Kernel Hilbert Space.

Definition 1 (see [Reproducing Kernel Hilbert Space]). Let $\mathcal{X}$ be a nonempty set and $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. Let $\langle \cdot, \cdot \rangle$ be a dot product in $\mathcal{H}$ which induces a norm as $\| f \| = \sqrt{\langle f, f \rangle}$. $\mathcal{H}$ is called an RKHS if there is a function $k : \mathcal{X} \times \mathcal{X}$ with the following properties:

(i) k has the reproducing property $\langle f, k(x, \cdot) \rangle = f(x)$ for all $f \in \mathcal{H}, x \in \mathcal{X}$;

(ii) k spans $\mathcal{H}$, that is, $\mathcal{H} = \overline{\operatorname{span}\{k(x, \cdot) \mid x \in \mathcal{X}\}}$, where $\overline{X}$ is the completion of the set X;

(iii) k is symmetric, that is, $k(x, y) = k(y, x)$.

Next, we introduce the Hyper Reproducing Kernel Hilbert Space.

Definition 2 (see [Hyper-Reproducing Kernel Hilbert Space]). Let $\mathcal{X}$ be a nonempty set and $\underline{\mathcal{X}} = \mathcal{X} \times \mathcal{X}$ be the Cartesian product. Let $\underline{\mathcal{H}}$ be the Hilbert space of functions $k : \underline{\mathcal{X}} \to \mathbb{R}$ with a dot product $\langle \cdot, \cdot \rangle$ and a norm $\|k\| = \sqrt{(\langle k, k \rangle)}$. $\underline{\mathcal{H}}$ is a Hyper Reproducing Kernel Hilbert Space if there is a hyperkernel $\underline{k} : \underline{X} \times \underline{X} \to \mathbb{R}$ with the following properties:

(i) $\underline{k}$ has the reproducing property $\langle k, \underline{k}(\underline{x}, \cdot) \rangle = k(\underline{x})$ for all $k \in \underline{\mathcal{H}}$;

(ii) $\underline{k}$ spans $\underline{H} = \overline{\operatorname{span}\{\underline{k}(\underline{x}, \cdot) \mid \underline{x} \in \underline{X}\}}$;

(iii) $\underline{k}(x, y, s, t) = \underline{k}(y, x, s, t)$ for all $x, y, s, t \in \mathcal{X}$.

Acknowledgments

The authors would like to thanks two anonymous referees by their useful comments and suggestions. Financial support from Grant S02EIA-07L01 is gratefully appreciated.

References

[1] Y. Hoshida, J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Subclass mapping: identifying common subtypes in independent disease data sets," *PLoS ONE*, vol. 2, no. 11, article e1195, pp. 1–8, 2007.

[2] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[3] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[4] Á. Blanco, M. Martín-Merino, and J. De Las Rivas, "Combining dissimilarity based classifiers for cancer prediction using gene expression profiles," *BMC Bioinformatics*, vol. 8, supplement 8, article S3, pp. 1–2, 2007.

[5] K. Tsuda, "Support vector classifier with asymmetric kernel function," in *Proceedings of the 7th European Symposium on Artificial Neural Networks (ESANN '99)*, pp. 183–188, Bruges, Belgium, April 1999.

[6] B. Schölkopf, J. Weston, E. Eskin, C. Leslie, and W. Stafford Noble, "A kernel approach for learning from almost orthogonal patterns," in *Proceedings of the 13th European Conference on Machine Learning (ECML '02)*, vol. 2430 of *Lecture Notes in Computer Science*, pp. 511–528, Springer, Helsinki, Finland, August 2002.

[7] N. Cristianini, J. Kandola, J. Elisseeff, and A. Shawe-Taylor, "On the kernel target alignment," *Journal of Machine Learning Research*, vol. 1, pp. 1–31, 2002.

[8] J. Kandola, J. Shawe-Taylor, and N. Cristianini, "Optimizing kernel alignment over combinations of kernels," Tech. Rep. NC-TR-02-121, NeuroCOLT, London, UK, 2002.

[9] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[10] C. S. Ong, A. J. Smola, and R. C. Williamson, "Learning the kernel with hyperkernels," *Journal of Machine Learning Research*, vol. 6, pp. 1043–1071, 2005.

[11] R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, Berlin, Germany, 2006.

[12] G. Wu, E. Y. Chang, and N. Panda, "Formulating distance functions via the kernel trick," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 703–709, Chicago, Ill, USA, August 2005.

[13] E. Pekalska, P. Paclick, and R. Duin, "A generalized kernel approach to dissimilarity-based classification," *Journal of Machine Learning Research*, vol. 2, pp. 175–211, 2001.

[14] S. Ramaswamy, P. Tamayo, R. Rifkin, et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 15149–15154, 2001.

[15] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.

[16] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11-12, no. 1–4, pp. 625–653, 1999.

[17] J. Löfberg, YALMIP, yet another LMI parser, 2002, http://control.ee.ethz.ch/~joloef/wiki/pmwiki.php.

[18] S. Fine and K. Scheinberg, "Efficient SVM training using low-rank kernel representations," *Journal of Machine Learning Research*, vol. 2, pp. 243–264, 2001.

[19] M. A. Shipp, K. N. Ross, P. Tamayo, et al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[20] K. J. Savage, S. Monti, J. L. Kutok, et al., "The molecular signature of mediastinal large B-cell lymphoma differs from that of other diffuse large B-cell lymphomas and shares features with classical Hodgkin lymphoma," *Blood*, vol. 102, no. 12, pp. 3871–3879, 2003.

[21] R. A. Irizarry, B. Hobbs, F. Collin, et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[22] M. West, C. Blanchette, H. Dressman, et al., "Predicting the clinical status of human breast cancer by using gene expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 20, pp. 11462–11467, 2001.

[23] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, et al., "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[24] I. B. Jeffery, D. G. Higgins, and A. C. Culhane, “Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data,” *BMC Bioinformatics*, vol. 7, article 359, pp. 1–16, 2006.

[25] S. Dudoit, J. Fridlyand, and T. P. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data,” *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–86, 2002.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 623853, 7 pages
doi:10.1155/2009/623853

Methodology Report

A Bayesian Approach to Multistage Fitting of the Variation of the Skeletal Age Features

Dong Hua,[1] Dechang Chen,[2] Fang Liu,[3] and Abdou Youssef[1]

[1] *Department of Computer Science, The George Washington University, 801 22nd Street NW, Washington, DC 20052, USA*
[2] *Division of Epidemiology and Biostatistics, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, USA*
[3] *Department of Computer Science, University of Texas-Pan American, 1201 W. University Drive, Edinburg, TX 78539, USA*

Correspondence should be addressed to Dechang Chen, dchen@usuhs.mil

Received 1 December 2008; Accepted 17 March 2009

Recommended by Zhenqiu Liu

Accurate assessment of skeletal maturity is important clinically. Skeletal age assessment is usually based on features encoded in ossification centers. Therefore, it is critical to design a mechanism to capture as much as possible characteristics of features. We have observed that given a feature, there exist stages of the skeletal age such that the variation pattern of the feature differs in these stages. Based on this observation, we propose a Bayesian cut fitting to describe features in response to the skeletal age. With our approach, appropriate positions for stage separation are determined automatically by a Bayesian approach, and a model is used to fit the variation of a feature within each stage. Our experimental results show that the proposed method surpasses the traditional fitting using only one line or one curve not only in the efficiency and accuracy of fitting but also in global and local feature characterization.

1. Introduction

Hand X-ray shown in Figure 1 is commonly used for skeletal age assessment in pediatric radiology. A discrepancy between skeletal maturity and the chronical age may indicate the presence of some abnormality in skeletal growth. This abnormality has been found to be related to various diseases such as endocrine disorders [1], metabolic/growth abnormalities [2], malformations and bone dysplasias [3], and gonadal dysgenesis [4]. Therefore, the assessment of skeletal maturity has become more and more important clinically. Clearly the accuracy in assessment is of the first concern.

Features encoded in ossification centers form the basis for assessment. If we know the exact characteristics of the features with regard to different stages of ages, we can do the best job on assessment. In reality, one needs a mechanism to capture such characteristics of features. Given data of a feature with respect to skeletal ages, a simple and common approach is to fit a line or a curve, which in turn is used for future prediction of new patients or assisting radiologists to understand the variation rules of the feature.

For instance, Figure 2(a) shows the variation of a ratio feature [5, 6] in vertical axis with regard to the increasing skeletal age along the horizontal axis from newborn to 19 year old boys. (More details on this ratio are provided in Section 3.2.) Here in the figure, a single line is used for fitting the values of the feature. Obviously, a line is not enough to capture the characteristic of the values of the feature. A quadratic curve, shown in Figure 2(c), does not do a good job either. Fitting a more complex curve does not seem to be a feasible approach. This is because sometimes there are available only a small amount of data which could restrict the learning of complex curves, and local properties (with respect to the time) of the feature are often lost when fitting a global complex curve, and thus leading to inaccurate future prediction.

In this paper, we propose to fit the variation of features of the skeleton age via a multistage fitting approach. With our approach, we divide the skeletal age axis into several stages or phases, and within each stage, a relative simple model (line or curve) is employed for the purpose of fitting. Usually, the variation of a feature does not follow a simple rule

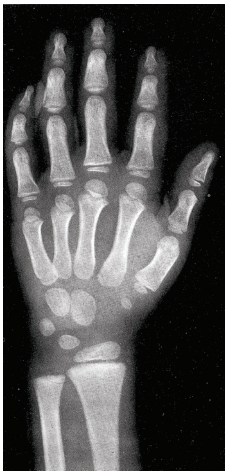


Figure 1: Hand X-ray used in skeletal age assessment.

when skeletal age increases. Instead, it often shows different variation patterns among different stages of age. As shown in Figures 2(b) and 2(d), multistage fitting not only can capture the entire pattern of feature variation but also carry the local properties regarding the skeletal age. A critical question is then, how does one determine the appropriate positions to separate the stages? The proposed *Bayesian cut* in this paper provides an answer via a Bayesian approach.

The rest of the paper is organized as follows. In Section 2, we describe our models for fitting, where the Bayesian cut is introduced. In Section 3, we present our experimental results on multi-stage fitting for artificial and real data. We conclude our paper in Section 4.

2. The Proposed Method

In this section, we first describe our proposed method for a simple case and then extend it to a general scenario.

Given a sequence of values $f_1, f_2, \dots, f_n$, which denotes the skeletal age f in an ascending order, consider the linear relationship between f and one feature y found in the hand X-ray (e.g., length of digit). Usually, such a linear relationship varies as the skeletal age increases. That is, one linear form established for one interval of the skeletal age may not hold for the next interval, where a different linear form should be used. The time where two linear forms differ is called a *change point*. Our model that takes into account linear relationships and change points is stated as follows:

$$\begin{aligned} y_i &= \beta_{11} + \beta_{12} f_i + \epsilon_{1i}, \quad i = 1, \dots, t_1\ (t_0 = 0), \\ y_i &= \beta_{21} + \beta_{22} f_i + \epsilon_{2i}, \quad i = t_1 + 1, \dots, t_2, \\ &\ \vdots \\ y_i &= \beta_{k1} + \beta_{k2} f_i + \epsilon_{ki}, \quad i = t_{k-1} + 1, \dots, t_k\ (t_k = n), \end{aligned} \tag{1}$$

where $t_1, \dots, t_{k-1}$ (correspondingly $f_1, \dots, f_{k-1}$) indicate the sequential change points, $t_j - t_{j-1} \geq 3$ $(j = 1, \dots, k)$, and ϵ_{ji} (for all i) are independent $N(0, \sigma_j^2)$ and ϵ_{ji} (for all i, j) are independent of each other. In the model, the parameters β_{j1}, β_{j2}, σ_j^2, t_j are all unknown, which will be estimated in light of the given data. The interval $[t_j - t_{j-1}]$ represents the jth stage or phase, denoted by ph_j. The main task here is to estimate the times t_j. Given the estimates of t_j, the linear forms and the associated parameters can be obtained through the traditional regression technique. We note that the requirement $t_j - t_{j-1} \geq 3\,(j = 1, \dots, k)$ is needed for estimation of the regression lines. When $k = 2$, the model will be reduced to the two-phase regression with a single change point in [7].

The above model that uses only one dependent variable f can be generalized to include multiple independent variables. This generalization leads to the following model:

$$\begin{aligned} y_i &= \vec{\beta}_1^T \mathbf{f_i} + \epsilon_{1i}, \quad i = 1, \dots, t_1\ (t_0 = 0), \\ y_i &= \vec{\beta}_2^T \mathbf{f_i} + \epsilon_{2i}, \quad i = t_1 + 1, \dots, t_2, \\ &\ \vdots \\ y_i &= \vec{\beta}_k^T \mathbf{f_i} + \epsilon_{ki}, \quad i = t_{k-1} + 1, \dots, t_k\ (t_k = n), \end{aligned} \tag{2}$$

where $\mathbf{f_i}$ is a p-dimensional vector of variables, $\vec{\beta}_j$ $(j = 1, \dots, k)$ is a p-dimensional vector of parameters, $t_j - t_{j-1} \geq p + 1$, and ϵ_{ji} are as the same as before. We refer p as the cardinality of the input vector $\mathbf{f_i}$, denoted by $C(\mathbf{f_i})$, and the number of sample points in ph_j as the cardinality of $[t_j - t_{j-1}]$, denoted by $C(ph_j)$. We note that though linear regression is used for each phase in model (2), this model certainly encompasses other nonlinear cases such as polynomial forms.

We now describe a Bayesian approach to estimate the change points. Denote $(\mathbf{f_{t_{j-1}+1}}, \dots, \mathbf{f_{t_j}})^T$ by F_j, $(F_1^T, \dots, F_k^T)^T$ by F, $(y_{t_{j-1}+1}, \dots, y_{t_j})^T$ by $\mathbf{y_j}$, $(\mathbf{y_1}^T, \dots, \mathbf{y_k}^T)^T$ by $\mathbf{y}$, and $(t_1, \dots, t_{k-1})$ by $\mathbf{t}$. For simplicity, we assume the noninformative or uniform prior for $\vec{\beta}_j$ $(j = 1, \dots, k)$, $\ln({\sigma_j}^2)$ and $\mathbf{t}$. Noninformative priors are used when information about parameters is completely unknown or when proper priors such as conjugate priors do not apply. (For a vigorous discussion on the choice of priors, see [8].) We can show the following main result (see the Appendix). Given the data $\mathbf{y}$ and the uniform prior for $\vec{\beta}_j$ $(j = 1, \dots, k)$, $\ln({\sigma_j}^2)$ and $\mathbf{t}$, where the number k is predetermined, the posterior probability that change points occur at $\mathbf{t}$ is

$$\begin{aligned} p(\mathbf{t} \mid \mathbf{y}) = {} & J 2^{(n-kp)/2} \prod_j \left| {F_j}^T F_j \right|^{-1/2} \\ & \times \Gamma\left(\frac{t_j - t_{j-1} - p}{2} \right) S_j^{-(t_j - t_{j-1} - p)/2}, \end{aligned} \tag{3}$$

where $J = \left(\sum_{\mathbf{t}} 2^{(n-kp)/2} \prod_j |F_j^T F_j|^{-1/2} \Gamma((t_j - t_{j-1} - p)/2) \times S_j^{-(t_j - t_{j-1} - p)/2} \right)^{-1}$, and $S_j = (\mathbf{y_j} - F_j \hat{\vec{\beta}}_j)^T (\mathbf{y_j} - F_j \hat{\vec{\beta}}_j)$ with $\hat{\vec{\beta}}_j = (F_j^T F_j)^{-1} F_j^T \mathbf{y_j}$ denoting the least-squares estimator of $\vec{\beta}_j$. Using this result, we estimate $\mathbf{t}$ by t^* at which $p(\mathbf{t} \mid \mathbf{y})$

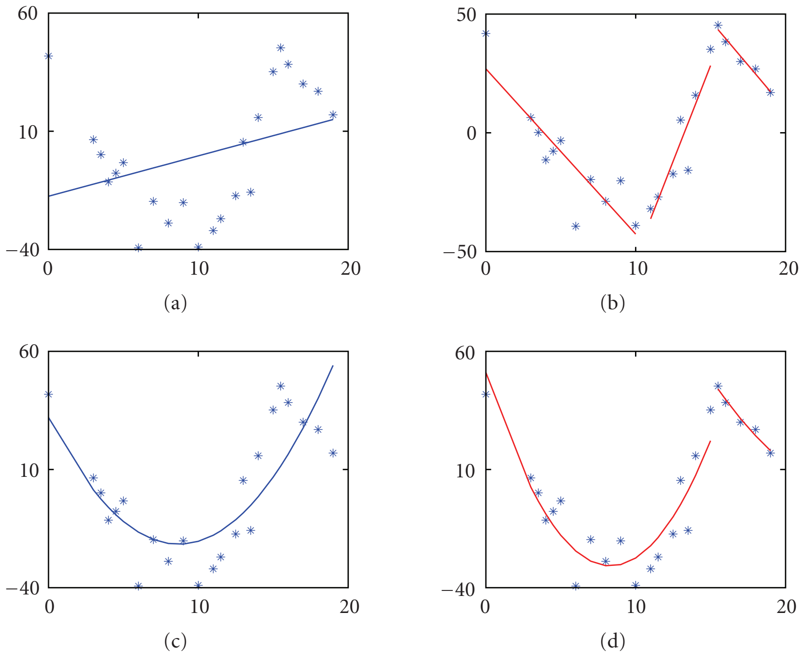


FIGURE 2: Examples of fitting the variation of the ratio feature. The horizontal axis represents the skeletal age and the vertical axis corresponds to the values of the feature.

TABLE 1: Models for testing the performance of the Bayesian cut.

m_1	$y_i = \beta_{j1} + \beta_{j2} f_i + \epsilon_{ji}$, $\mathbf{t} = (t_1, \dots, t_{k-1})$
m_2	$y_i = \beta_{j1} + \beta_{j2} f_i + \beta_{j3} f_i^2 + \epsilon_{ji}$, $\mathbf{t} = (t_1, \dots, t_{k-1})$
m_3	$y_i = \beta_{j1} + \beta_{j2} f_i + \beta_{j3} f_i^2 + \beta_{j4} f_i^3 + \epsilon_{ji}$, $\mathbf{t} = (t_1, \dots, t_{k-1})$
m_4	$y_i = \beta_{j1} + \beta_{j2} f_i + \beta_{j3} f_i^2 + \beta_{j4} f_i^3 + \beta_{j5} f_i^4 + \epsilon_{ji}$, $\mathbf{t} = (t_1, \dots, t_{k-1})$
m_5	$y_i = \beta_{j1} + \beta_{j2} f_i + \beta_{j3} f_i^2 + \beta_{j4} f_i^3 + \beta_{j5} f_i^4 + \beta_{j6} f_i^5 + \epsilon_{ji}$, $\mathbf{t} = (t_1, \dots, t_{k-1})$

TABLE 2: Experimental setting.

β_{ji}	$(-5.0, 5.0)$
ϵ_{ji}	$\sim N(0, \sigma_j^2),\ \sigma_j^2 \in (0, 5^{C(\mathbf{f}_i)-1})$
k	2, 3, 4
$C(ph_j)$	$(C(\mathbf{f}_i)+1), \dots, (C(\mathbf{f}_i)+1)+s$
scale	$1, \dots, 10$
t_0	0
t_j	$t_{j-1} + C(ph_{j-1})$
f_i	$1, \dots, t_k$

has its maximum, that is, $t^* = \arg\max_{\mathbf{t}} p(\mathbf{t} \mid \mathbf{y})$. We call t^* the *Bayesian cut*, and the value $2^{(n-kp)/2} \prod_j |F_j^T F_j|^{-1/2} \Gamma((t_j - t_{j-1} - p)/2) S_j^{-(t_j - t_{j-1} - p)/2}$ the *proportional posterior* (pp).

3. Experiments

In this section, we perform the Bayesian cut on two data sets: one is synthesized and the other is real. We use the synthesized data for performance evaluation in terms of recovery of changing points. The real data are used to discover the Bayesian cut and describe the feature in a multistage way which has more accurate prediction of the skeletal age compared with fitting by a single line or curve. Both linear and nonlinear regression are used for comparison. For convenience, we call the fitting with a single line or curve the *single fitting* and the fitting with the Bayesian cut the *Bayesian cut fitting*.

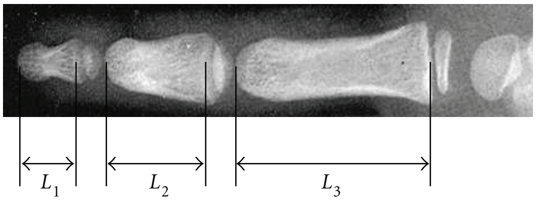


FIGURE 3: Illustration Of L_1, L_2 and L_3.

3.1. Synthesized Data. We consider five cases or models describing the relationship between the dependent and independent variables. These are shown in Table 1 where the input vector $\mathbf{f_i}$ for models m_1, m_2, m_3, m_4, and m_5 is $(1, f_i)^T$, $(1, f_i, f_i^2)^T$, $(1, f_i, f_i^2, f_i^3)^T$, $(1, f_i, f_i^2, f_i^3, f_i^4)^T$, and $(1, f_i, f_i^2, f_i^3, f_i^4, f_i^5)^T$, respectively. The data are generated according to the setting given in Table 2. Specifically, β_{ji} is randomly chosen from $(-5.0, 5.0)$. ϵ_{ji} is generated from a normal distribution with mean 0 and variance σ_j^2 randomly selected from $(0, 5^{C(\mathbf{f}_i)-1})$. The number of sample points of

Table 3: AD scores for models in Table 1.

k	s	m_1	m_2	m_3	m_4	m_5
2	1	0.280	0.340	0.320	0.080	0.180
	2	0.300	0.460	0.360	0.200	0.100
	3	0.260	0.400	0.320	0.100	0.100
	4	0.640	0.380	0.260	0.180	0.180
	5	0.480	0.680	0.480	0.100	0.060
	6	0.380	0.300	0.560	0.220	0.100
	7	0.540	0.520	0.340	0.280	0.100
	8	0.900	0.520	0.440	0.120	0.020
	9	0.740	0.340	0.080	0.040	0.020
	10	0.740	0.720	0.160	0.200	0.020
3	1	0.230	0.360	0.210	0.240	0.090
	2	0.440	0.390	0.190	0.080	0.060
	3	0.590	0.340	0.210	0.220	0.060
	4	0.820	0.590	0.260	0.060	0.010
	5	0.970	0.690	0.530	0.020	0.090
	6	0.670	0.580	0.120	0.060	0.070
	7	1.220	0.750	0.160	0.080	0.190
	8	1.260	0.680	0.650	0.040	0.030
	9	1.210	0.860	0.370	0.380	0.010
	10	1.340	0.360	0.680	0.020	0.020
4	1	0.333	0.300	0.133	0.040	0.053
	2	0.440	0.433	0.227	0.060	0.033
	3	0.867	0.480	0.113	0.080	0.033
	4	0.780	0.513	0.093	0.080	0.133
	5	1.020	0.887	0.453	0.133	0.173
	6	1.360	0.760	0.193	0.093	0.180
	7	1.007	0.593	0.353	0.047	0.040
	8	0.727	0.587	0.453	0.093	0.113
	9	1.080	1.240	0.867	0.360	0.087
	10	1.213	0.873	0.333	0.120	0.140

the jth phase $C(ph_j)$ is randomly selected from the set $\{(C(\mathbf{f_i}) + 1), \ldots, (C(\mathbf{f_i}) + 1) + s\}$, where s is predetermined. f_i takes the value of i for $i = 1, 2, \ldots, t_k$. Note that we use a variable bound for σ_j^2 for taking into account the influence of the highest degree of the polynomial. Also, we use the variable number of sample points for each phase by introducing unbalance and scalability factors such that the performance evaluation will be more objective.To present a quantity on the performance of the Bayesian cut, we use the metric *absolute deviation* (AD), defined as

$$\mathrm{AD} = \frac{\sum_j \left| t_j^* - t_j \right|}{k-1}, \qquad j = 1, \ldots, k-1, \tag{4}$$

where t_j^* represents the jth element of $\mathbf{t}^*$ (the Bayesian cut). Intuitively, the smaller AD is, the closer is the Bayesian cut $\mathbf{t}^*$ to the true change points $\mathbf{t}$.

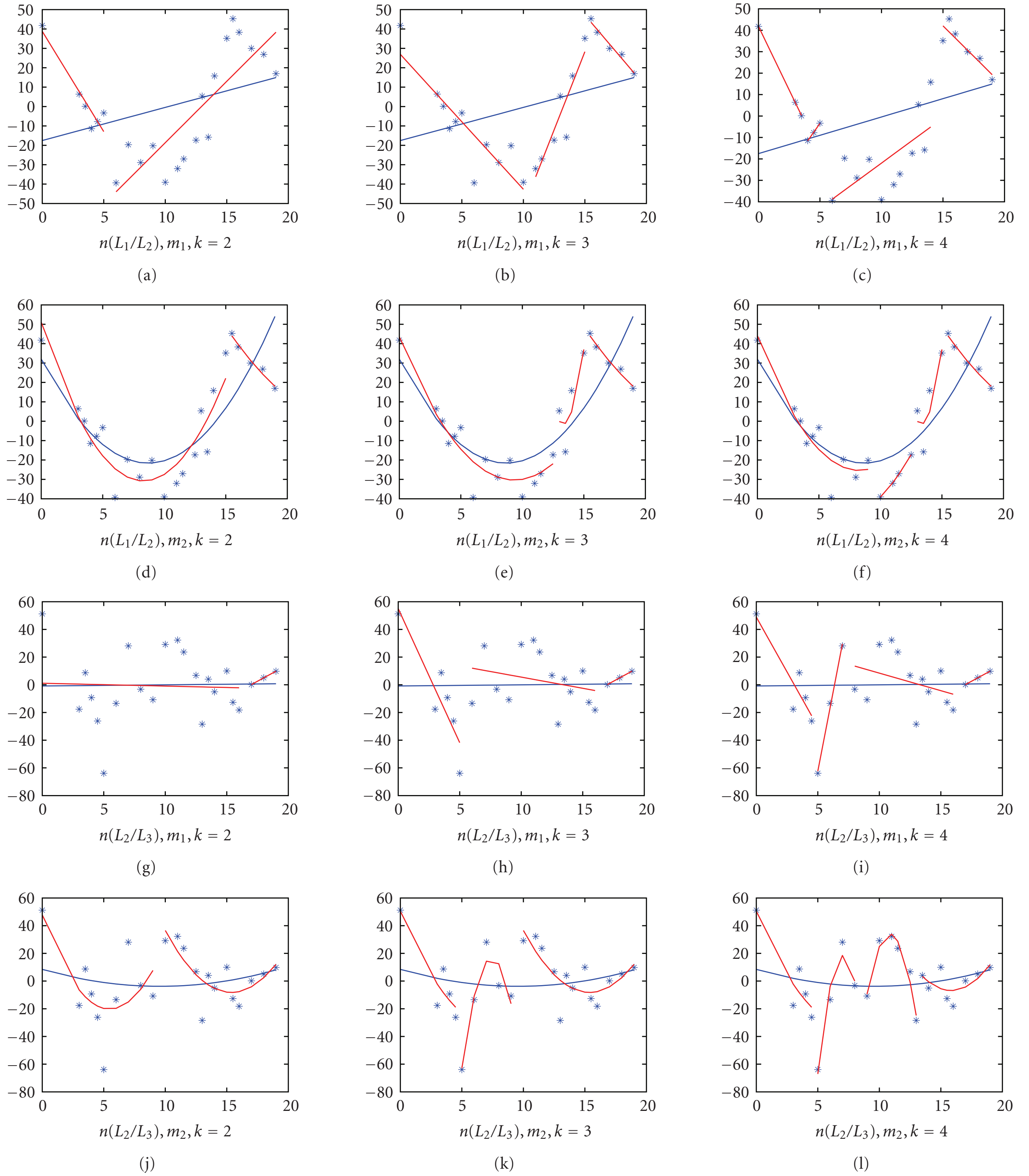
Table 3 shows the AD values. They are obtained by ranging k from 2 to 4 and s from 1 to 10. For given k, s, and a given model, 50 trials are performed to generate data, leading to 50 datasets $\{(F, \mathbf{y})\}$. We find the Bayesian cut $\mathbf{t}^*$ for each $(F, \mathbf{y})$ and a given model. The final AD score is obtained by averaging the 50 runs.

Table 4: Some features of the skeletal age.

Age (yr)	L_1/L_2	L_2/L_3	$n(L_1/L_2)$	$n(L_2/L_3)$
0	0.6795	0.7016	41.8212	51.1987
3	0.6307	0.5853	6.4071	−17.6281
3.5	0.6220	0.6298	0.1020	8.6933
4.0	0.6060	0.5993	−11.4491	−9.3140
4.5	0.6111	0.5708	−7.7721	−26.1616
5.0	0.6172	0.5070	−3.3303	−63.8970
6.0	0.5675	0.5924	−39.3612	−13.4245
7.0	0.5947	0.6626	−19.6939	28.0937
8.0	0.5820	0.6097	−28.9032	−3.1878
9.0	0.5939	0.5968	−20.2149	−10.7828
10.0	0.5680	0.6643	−39.0383	29.1323
11.0	0.5776	0.6696	−32.0541	32.2560
11.5	0.5845	0.6550	−27.0602	23.6424
12.5	0.5979	0.6266	−17.3472	6.8003
13.0	0.6292	0.5670	5.3295	−28.4227
13.5	0.6000	0.6219	−15.8024	4.0436
14.0	0.6436	0.6065	15.7982	−5.0842
15.0	0.6703	0.6319	35.1558	9.9431
15.5	0.6843	0.5937	45.2891	−12.6564
16.0	0.6746	0.5843	38.2966	−18.2156
17.0	0.6632	0.6153	30.0081	0.1412
18.0	0.6589	0.6236	26.8770	5.0546
19.0	0.6452	0.6316	16.9420	9.7754

Our findings can be summarized as follows. Regardless of linear or nonlinear regression, the Bayesian cut performs well with low AD scores. Introducing the unbalance and scalability factors does not deteriorate the performance of the Bayesian cut significantly. The Bayesian cut scales well when the number of change points increases.

3.2. Real Data. In this part, we apply the Bayesian cut fitting to some real data from our database shown in Table 4. This table describes feature values with regard to the increasing skeletal age that ranges from newborn to 19-year-old boys (shown in column 1) labeled by radiology experts. In order to obtain features independent of the size and the length of digits, two ratio features are used according to the paper [5]. One is L_1/L_2, the ratio of the length of distal phalanx L_1 to that of middle phalanx L_2 of the middle digit, and the other is L_2/L_3, the ratio of the length of middle phalanx L_2 to that of proximal phalanx L_3. See Figure 3 for illustration of L_1, L_2, and L_3. These two features correspond to columns 2 and 3 which are generated in the light of the algorithm in [6]. Columns 4 and 5 represent normalized values of L_1/L_2 and L_2/L_3, respectively. This normalization is done according to $(x - \mu)/\sigma$, where μ is the expectation of x and σ is the variance. In our experiments, only normalized values are used. Figure 4 shows some of the Bayesian cut fitting, where features $n(L_1/L_2)$ and $n(L_2/L_3)$ are used, models describing the relationship between the feature and the skeletal age are



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 4: Illustration of the Bayesian cut fitting applied to the real data on features of the skeletal age.

m_1 and m_2 from Table 1, and k takes values of 2, 3, and 4. In Figure 4, the horizontal axis represents the age and the horizontal axis indicates the feature. For model m_1, the blue straight line across the entire age range is from the single (line) fitting. For model m_2, the blue curve across the entire age range is from the single (quadratic) fitting. All red (broken) lines are from the Bayesian cut fitting.

4. Conlcusion

In this paper, we propose the Bayesian cut fitting to describe features in response to the skeletal age. In the semantic space derived by our approach, the axis of skeletal age is divided into meaningful stages, within each of which the variation pattern of a feature is consistent so that a traditional regression technique can apply to model the relationship between the skeletal age and the feature. Our approach is inspired by the observation that the variation pattern of a feature can differ in different periods of the skeletal age. A critical issue is to determine the times or change points when the variation pattern of a feature changes. This is handled by the Bayesian cut proposed in this paper. Simulations have been used to demonstrate the efficiency of the Bayesian cut fitting in terms of recovery of change points. The experiments on real data show that given a type of relationship (e.g., linear or quadratic) between the skeletal age and a feature, the Bayesian cut fitting surpasses the traditional single fitting when the consistency of the variation pattern (over the entire skeletal age range) of the feature is suspected. One major issue which is not addressed in this paper is the determination of k, the number of stages. Selection of k depends on the given data and the practical need. We leave this as our future research work.

Appendix

A. Derivation of (3)

Proof. According to the Pythagorean theorem, we have the following likelihood

$$\begin{aligned} l\left(\vec{\beta}_j, \sigma_j^2 \mid \mathbf{y}\right) &\propto \frac{1}{\left(\sigma_j^2\right)^{(t_j-t_{j-1})/2}} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma_j{}^2}\left[\left(y_{t_{j-1}+1} - \vec{\beta}_j^T \mathbf{f}_{\mathbf{t}_{j-1}+1}\right)^2 \right.\right. \\ &\quad \left.\left. + \cdots + \left(y_{t_j} - \vec{\beta}_j^T \mathbf{f}_{\mathbf{t}_\mathbf{j}}\right)^2\right]\right\} \\ &\propto \frac{1}{\left(\sigma_j^2\right)^{(t_j-t_{j-1})/2}} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma_j{}^2}\left[S_j + \left(\vec{\beta}_j - \hat{\vec{\beta}}_j\right)^T \right.\right. \\ &\quad \left.\left. \times F_j^T F_j\left(\vec{\beta}_j - \hat{\vec{\beta}}_j\right)\right]\right\}, \end{aligned} \tag{A.1}$$

where $S_j = (\mathbf{y_j} - F_j \hat{\vec{\beta}}_j)^T(\mathbf{y_j} - F_j \hat{\vec{\beta}}_j)$ and $\hat{\vec{\beta}}_j = (F_j^T F_j)^{-1} F_j^T \mathbf{y_j}$. Since ϵ_{ji} are independent of each other, the likelihood function of $\vec{\beta}_1, \ldots, \vec{\beta}_k$, $\sigma_1^2, \ldots, \sigma_k^2, \mathbf{t}$ is then

$$\begin{aligned} & l\left(\vec{\beta}_1, \ldots, \vec{\beta}_k, \sigma_1^2, \ldots, \sigma_k^2, \mathbf{t} \mid \mathbf{y}\right) \\ &\quad \propto \prod_j \frac{1}{\left(\sigma_j^2\right)^{(t_j-t_{j-1})/2}} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma_j^2}\left[S_j + \left(\vec{\beta}_j - \hat{\vec{\beta}}_j\right)^T F_j^T F_j\left(\vec{\beta}_j - \hat{\vec{\beta}}_j\right)\right]\right\}. \end{aligned} \tag{A.2}$$

Due to the assumption of the uniform prior for $\vec{\beta}_j$, $\ln(\sigma_j^2)$ and $\mathbf{t}$, we have

$$p\left(\vec{\beta}_1, \ldots, \vec{\beta}_k, \sigma_1^2, \ldots, \sigma_k^2, \mathbf{t}\right) \propto \frac{1}{\sigma_1^2 \cdots \sigma_k^2}. \tag{A.3}$$

Using (A.2) and (A.3), we have

$$\begin{aligned} & p(\mathbf{t} \mid \mathbf{y}) \\ &\propto \int_{\vec{\beta}_1} \cdots \int_{\vec{\beta}_k} \int_{\sigma_1^2} \cdots \int_{\sigma_k^2} \prod_j \frac{1}{\left(\sigma_j^2\right)^{(t_j-t_{j-1})/2+1}} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma_j{}^2}\left[S_j + \left(\vec{\beta}_j - \hat{\vec{\beta}}_j\right)^T \right.\right. \\ &\quad \left.\left. \times F_j^T F_j\left(\vec{\beta}_j - \hat{\vec{\beta}}_j\right)\right]\right\} \\ &\quad \times d\vec{\beta}_1 \cdots d\vec{\beta}_k d\sigma_1^2 \cdots d\sigma_k^2. \end{aligned} \tag{A.4}$$

Note that

$$\begin{aligned} & \int_{\vec{\beta}_j} \frac{1}{\left(\sigma_j^2\right)^{(t_j-t_{j-1})/2+1}} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma_j{}^2}\left[S_j + \left(\vec{\beta}_j - \hat{\vec{\beta}}_j\right)^T F_j^T F_j\left(\vec{\beta}_j - \hat{\vec{\beta}}_j\right)\right]\right\} d\vec{\beta}_j \\ &\quad = \frac{\exp\left(-S_j/2\sigma_j^2\right)}{\left(\sigma_j^2\right)^{(t_j-t_{j-1})/2+1}} (2\pi)^{p/2}\left(2\sigma_j^2\right)^{p/2}\left|F_j^T F_j\right|^{-1/2}. \end{aligned} \tag{A.5}$$

This equation exploits the fact

$$\int \exp\left\{(\mathbf{x} - \vec{\mu})^T \Sigma^{-1}(\mathbf{x} - \vec{\mu})\right\} d\mathbf{x} = (2\pi)^{p/2}|\Sigma|^{1/2}, \tag{A.6}$$

from the normal density for the p-dimensional random vector X

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{(\mathbf{x} - \vec{\mu})^T \Sigma^{-1}(\mathbf{x} - \vec{\mu})\right\} d\mathbf{x}, \tag{A.7}$$

where $\vec{\mu}$ is the expected value of X and Σ is the variance-covariance matrix of X.

Substituting (A.5) into (A.4), we have

$$p(\mathbf{t} \mid \mathbf{y}) \propto \prod_j \left|F_j^T F_j\right|^{-1/2} \int_{\sigma_j^2} \frac{\exp\left(-S_j/2\sigma_j^2\right)}{\left(\sigma_j^2\right)^{(t_j-t_{j-1}-p)/2+1}} d\sigma_j^2. \tag{A.8}$$

In addition, we have

$$\int \exp\left(-\frac{a}{2x}\right) x^{-m/2-1} dx = 2^{m/2}\Gamma\left(\frac{m}{2}\right) a^{-m/2}, \tag{A.9}$$

from the probability density function of $X = aU$

$$f(x) = 2^{-m/2}\left[\Gamma\left(\frac{m}{2}\right)\right]^{-1} a^{m/2}x^{-m/2-1}\exp\left(-\frac{a}{2x}\right), \quad \text{(A.10)}$$

where the constant $a > 0$ and $U^{-1} \sim \chi^2_m$.

By applying (A.9) to (A.8), we get

$$\begin{aligned} p(\mathbf{t} \mid \mathbf{y}) \propto J2^{(n-kp)/2}\prod_j \left|F_j^T F_j\right|^{-1/2} \\ \times \Gamma\left(\frac{t_j - t_{j-1} - p}{2}\right) S_j^{-(t_j-t_{j-1}-p)/2}, \end{aligned} \quad \text{(A.11)}$$

where $J = (\sum_{\mathbf{t}} 2^{(n-kp)/2}\prod_j |F_j^T F_j|^{-1/2}\Gamma((t_j - t_{j-1} - p)/2) \times S_j^{-(t_j-t_{j-1}-p)/2})^{-1}$. This completes the proof. □

Acknowledgments

Dechang Chen was partially supported by the National Science Foundation grant CCF-0729080.

References

[1] D. B. Darling, *Radiography of Infants and Children*, chapter 6, Charles C. Thomas, Springfield, Ill, USA, 1st edition, 1979.

[2] A. K. Poznanski, S. M. Garn, J. M. Nagy, and J. C. Gall Jr., "Metacarpophalangeal pattern profiles in the evaluation of skeletal malformations," *Radiology*, vol. 104, no. 1, pp. 1–11, 1972.

[3] D. R. Kirks, *Practical Pediatric Imaging: Diagnostic Radiology of Infants and Children*, chapter 6, Little, Brown, Boston, Mass, USA, 1st edition, 1984.

[4] J. Kosowicz, "The roentgen appearance of the hand and wrist in gonadal dysgenesis," *The American Journal of Roentgenology, Radium Therapy and Nuclear Medicine*, vol. 93, pp. 354–361, 1965.

[5] E. Pietka, M. F. McNitt-Gray, M. L. Kuo, and H. K. Huang, "Computer-assisted phalangeal analysis in skeletal age assessment," *IEEE Transactions on Medical Imaging*, vol. 10, no. 4, pp. 616–620, 1991.

[6] E. Pietka, A. Gertych, S. Pospiech, F. Cao, H. K. Huang, and V. Gilsanz, "Computer-assisted bone age assessment: image preprocessing and epiphyseal/metaphyseal ROI extraction," *IEEE Transactions on Medical Imaging*, vol. 20, no. 8, pp. 715–729, 2001.

[7] D. Chen, M. Fries, and J. M. Lyon, "A statistical method of detecting bioremediation," *Journal of Data Science*, vol. 1, no. 1, pp. 27–41, 2003.

[8] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, John Wiley & Sons, New York, NY, USA, 1992.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 707580, 11 pages
doi:10.1155/2009/707580

Research Article

Integrating Multiple Microarray Data for Cancer Pathway Analysis Using Bootstrapping K-S Test

Bing Han,[1] Xue-Wen Chen,[1] Xinkun Wang,[2] and Elias K. Michaelis[2]

[1] *Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA*
[2] *Higuchi Biosciences Center, University of Kansas, 2099 Constant Avenue, Lawrence, KS 66047, USA*

Correspondence should be addressed to Xue-Wen Chen, xwchen@ku.edu

Received 14 January 2009; Accepted 4 March 2009

Recommended by Dechang Chen

Previous applications of microarray technology for cancer research have mostly focused on identifying genes that are differentially expressed between a particular cancer and normal cells. In a biological system, genes perform different molecular functions and regulate various biological processes via interactions with other genes thus forming a variety of complex networks. Therefore, it is critical to understand the relationship (e.g., interactions) between genes across different types of cancer in order to gain insights into the molecular mechanisms of cancer. Here we propose an integrative method based on the bootstrapping Kolmogorov-Smirnov test and a large set of microarray data produced with various types of cancer to discover common molecular changes in cells from normal state to cancerous state. We evaluate our method using three key pathways related to cancer and demonstrate that it is capable of finding meaningful alterations in gene relations.

1. Introduction

Microarray technology, monitoring mRNA abundance of tens of thousands of genes simultaneously, provides an efficient tool to characterize a cell at the molecular level. It has been applied to a variety of research areas, ranging from biomarker detection [1, 2] to gene regulatory networks [3–5] and cancer classification [6–8]. When applied to cancer research, microarray technology typically measures gene expressions of cancer and normal tissues or different types of cancer. One important area in microarray-based cancer research is to identify genes that are differentially expressed between cancerous and normal cells and to discover diagnostic and prognostic signatures in order to predict therapeutic responses. Over the years, many statistical methods for the identification of differentially expressed genes have been developed, and most of them focused on the expression analysis of individual genes [9–15]. However, the simple list of individual differentially expressed genes can only tell us which genes are altered by biological differences between different cell types and/or states. It cannot explain the reasons for the significant alterations in gene expression levels and the effects of such changes on other genes' activities. It is well known that in a biological system genes interact with each other forming various biological pathways in order to carry out a multitude of biological processes. To better understand the roles of these differentially expressed genes and their interactions in a complex biological system, a comprehensive pathway analysis is needed. Since the identification of biological pathways is significantly influenced by those differentially expressed genes from different datasets or different statistical methods [16, 17], we reason here that an integration of multiple cancer microarray datasets and identification of the most common pathways from these data would reveal key relationships between crucial genes in carcinogenesis. Our focus on the interactions and pathways of cancer-related genes is important since changes in gene relations and key pathways are more relevant to carcinogenesis than individual genes alone.

Several statistical methods have been proposed for the analysis of differential gene coexpression patterns. Li [18] observed differences of gene coexpression patterns in different cellular states and attributed these changes in gene coexpression patterns to some third set of influential genes.

Lai et al. [19] proposed a similar method to identify differential gene-gene coexpression patterns in cells from normal state to cancerous state. However, these methods often perform the analyses on one single microarray dataset and typically generate unreliable results; the results from different microarray datasets and various statistical methods could hardly overlap using these methods [20, 21]. Therefore, the confidence level for discoveries based on these methods is low. Furthermore, these methods fail to grasp the common molecular changes in cells transitioning from a normal state to the cancerous state. Choi et al. [22] introduced a model to find differential gene coexpression patterns related to cancer by combining independent datasets for different cancers. They used a model similar to the t-test, which only considered the mean and variance of two groups of samples. It is well known that traditional t-test has two disadvantages for microarray data analysis: first, it assumes that the datasets under analysis have a normal distribution, which is usually violated in microarray datasets; second, if the number of genes is large and the number of samples is small, some of the standard deviations will be extremely small, and therefore the test statistics will be very high, which may lead to a significant bias. Nonparametric statistical test methods, such as the K-S test, require fewer assumptions for the data and may be preferred, especially, when the number of samples is small.

In this paper, we propose a novel method to detect the differentially changed gene relations in cancer versus normal tissues. We collect 36 datasets across different microarray platforms and from various types of cancer. These 36 datasets contain both normal and tumor samples, which can subsequently yield two Pearson correlation coefficient vectors for every gene pair, one for normal samples and the other for tumor samples. We then perform a bootstrapping K-S test to identify some differentially changed gene relations. Finally we verify our results with three key pathways related to cancer and demonstrate that our method can find some meaningful alterations of gene relations.

2. Materials and Methods

2.1. Microarray Datasets. We collected 36 microarray datasets from NCBI (Gene Expression Omnibus GEO) [23]. As shown in Table 1, these microarray datasets contain both normal and tumor samples across 21 different types of cancer, and their platforms come from one of the three platforms: GPL570 (Affymetrix GeneChip Human Genome U133 Plus 2.0 Array), GPL96 (Affymetrix GeneChip Human Genome U133 Array Set HG-U133A), and GPL91 (Affymetrix GeneChip Human Genome U95 Version Set HG-U95A). We divided every dataset into two expression data matrices: one matrix includes all normal samples, and the other includes all tumor samples. To integrate multiple microarray datasets across different platforms, we mapped each probe in different platforms to a unique Entrez Gene ID or a unique UniGene symbol. For genes with more than one probe in one platform, we chose the probe with the highest mean expression value.

2.2. Cancer-Associated Pathways and Extended Gene Networks. We applied our method to analyze three cancer-associated pathways. These pathways are related to three common traits in most and perhaps all types of human cancer: self-sufficiency in growth signals, insensitivity to antigrowth signals, and evading programmed cell death (apoptosis) [24]. In fact, Hanahan and Weinberg have already identified some signaling pathways to demonstrate the capabilities cancer cells acquire during tumor development in [24]. We extended these signaling pathways to three relatively complete and larger cancer-associated pathways (antigrowth signaling, apoptosis, and growth signaling pathways) from the cell cycle pathway, the apoptosis pathway and the MAPK pathway in KEGG [25]. We used these three pathways (i.e., cell cycle, apoptosis, and MAPK pathways) as our seeds and the genes in these pathways as our seed genes. Next we constructed three gene networks corresponding to the three cancer-associated pathways from HPRD (Human Proteins Reference Database, http://www.hprd.org/) and TRANSFAC [26] based on seed genes and their interacting partners. We downloaded the protein-protein interaction (PPI) data released by HPRD on September 1, 2007. This PPI dataset contains 37107 human binary protein-protein interactions whose supporting experiments are indicated as in vivo, in vitro, or yeast two-hybrid. We also collected 1042 transcription factor-target gene relations on human species from TRANSFAC. So our gene networks included seed genes, protein interaction partners, and transcription factors (TFs) of seed genes or target genes for which seed genes served as their TFs.

2.3. Detecting Differential Relations by Bootstrapping K-S Test. We used the Kolmogorov-Smirnov test (K-S test) to determine whether the distributions of values in two datasets differed significantly. The two-sample K-S test is the most useful for comparing two samples because it is nonparametric and distribution-free [27]. The null hypothesis for this test is that two datasets are drawn from the same distribution. The alternative hypothesis is that they are drawn from different distributions.

For n i.i.d samples $X_1, \ldots, X_n$ with some unknown distribution, we can define an empirical distribution function by

$$S_n(x) = \begin{cases} 0, & \text{if } x < X_{(1)}, \\ \dfrac{k}{n}, & \text{if } X_{(k)} \le x < X_{(k+1)} \quad \text{for } k = 1, 2, \ldots, n-1, \\ 1, & \text{if } x \ge X_{(n)}, \end{cases} \tag{1}$$

where $X_1, \ldots, X_n$ are ordered from the smallest to the largest value. The Kolmogorov-Smirnov statistic for a given function $S(x)$ is

$$D_n = \max_{\mathrm{x}} |S_n(x) - S(x)|. \tag{2}$$

D_n will converge to 0 if the sample comes from distribution $S(x)$ [27]. Moreover, the cumulative distribution function of

Table 1: List of 36 microarray datasets.

Series ID in GEO	Cancer type	Numbers of normal samples	Numbers of tumor samples	Numbers of genes	Platform ID in GEO
GSE3744	Breast cancer	7	40	54681	GPL570
GSE5764	Breast cancer	20	10	54681	GPL570
GSE7904	Breast cancer	19	43	54681	GPL570
GSE3678	Thyroid cancer	7	7	54681	GPL570
GSE3467	Thyroid cancer	9	9	54681	GPL570
GSE8977	Breast cancer	15	7	54681	GPL570
GSE8671	Colorectal cancer	32	32	54681	GPL570
GSE4290	Glioma	23	157	54681	GPL570
GSE4183	Colorectal cancer	8	30	54681	GPL570
GSE4107	Colorectal cancer	10	12	54681	GPL570
GSE8514	Aldosterone-producing adenoma	5	10	54681	GPL570
GSE6791	Cervical cancer	8	20	54681	GPL570
GSE6791	Head and neck cancer	18	38	54681	GPL570
GSE6338	Lymphoma	20	40	54681	GPL570
GSE5563	Vulvar intraepithelial neoplasia	9	9	54681	GPL570
GSE6004	Thyroid Cancer	4	14	54681	GPL570
GSE2549	Malignant pleural mesothelioma	10	44	22283	GPL96
GSE781	Kidney cancer	9	8	22283	GPL96
GSE7670	Lung cancer	27	27	22283	GPL96
GSE6344	Kidney cancer	10	10	22283	GPL96
GSE1542	Pancreatic ductal carcinoma	25	24	22283	GPL96
GSE6883	Breast cancer	6	6	22283	GPL96
GSE2724	Uterine fibroid	11	7	22283	GPL96
GSE2503	Skin cancer	6	5	22283	GPL96
GSE3268	Lung cancer	5	5	22283	GPL96
GSE9476	Acute myeloid leukemia	38	26	22283	GPL96
GSE6008	Ovarian tumor	4	99	22283	GPL96
GSE6477	Multiple myeloma	12	150	22283	GPL96
GSE4115	Lung Cancer	90	97	22283	GPL96
GSE3167	Bladder cancer	14	46	22283	GPL96
GSE2514	Pulmonary adenocarcinoma	19	20	12651	GPL91
GSE6631	Head and neck cancer	22	22	12651	GPL91
GSE6604	Prostate tumor	18	25	12651	GPL91
GSE6605					
GSE6606	Prostate tumor	63	65	12651	GPL91
GSE6608					
GSE2379	Head and neck cancer	4	34	12651	GPL91
GSE1987	Lung Cancer	9	28	12651	GPL91

Kolmogorov distribution is

$$K(x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty}e^{-(2i-1)^2\,\pi^2/(8x^2)}. \tag{3}$$

It is easy to prove that $\sqrt{n}D_n = \sqrt{n}\max_x |S_n(x) - S(x)|$ will converge to the Kolmogorov distribution [27]. Therefore if $\sqrt{n}D_n > K_\alpha = \Pr(K \leq K_\alpha) = 1 - \alpha$, the null hypothesis for the Kolmogorov-Smirnov test will be rejected at level α.

For the case of determining whether the distributions of two data vectors differ significantly, the Kolmogorov-Smirnov statistic is

$$D_{n,m} = \max_x |S_n(x) - S_m(x)|, \tag{4}$$

and the null hypothesis will be rejected at level α if

$$\sqrt{\frac{nm}{n+m}}D_{n,m} > K_{\alpha}. \tag{5}$$

The P-value from the K-S test can measure the confidence of the comparison results against the null hypothesis. Obviously, the smaller the P-value, the more confident we are of rejecting the null hypothesis.

Assume that we have n microarray datasets and a list of m genes, we denote the expression data matrix for normal samples as

$$N^k = \begin{pmatrix} X_{11}^k & X_{12}^k & \cdots & X_{1p}^k \\ X_{21}^k & X_{22}^k & \cdots & X_{2p}^k \\ \cdot & \cdot & \cdot & \cdot \\ X_{m1}^k & X_{m2}^k & \cdots & X_{mp}^k \end{pmatrix} \quad k = 1, \ldots, n, \tag{6}$$

and the expression data matrix for tumor samples as

$$T^l = \begin{pmatrix} Y_{11}^l & Y_{12}^l & \cdots & Y_{1q}^l \\ Y_{21}^l & Y_{22}^l & \cdots & Y_{2q}^l \\ \cdot & \cdot & \cdot & \cdot \\ Y_{m1}^l & Y_{m2}^l & \cdots & Y_{mq}^l \end{pmatrix} \quad l = 1, \ldots, n, \tag{7}$$

where $p(k)$ is the number of normal samples in the kth dataset, and $q(l)$ is the number of tumor samples in the lth dataset.

For these two types of expression data matrices, each row represents one gene, and each column represents one sample. The correlation coefficient for gene i and gene j from the kth normal sample can be calculated by

$$\mathrm{NPC}_{ij}^k = \frac{\sum_{a=1}^{p}\left(X_{ia}^k - \overline{X}_i^k\right)\left(X_{ja}^k - \overline{X}_j^k\right)}{\sqrt{\sum_{a=1}^{p}\left(X_{ia}^k - \overline{X}_i^k\right)^2}\sqrt{\sum_{a=1}^{p}\left(X_{ja}^k - \overline{X}_j^k\right)^2}}, \tag{8}$$

where $\overline{X}_i^k$ is the average value of expression levels for gene i. The correlation coefficient for every gene pair from tumor samples can be calculated similarly.

We use the bootstrapping K-S test to detect some gene relations with different PC (Pearson coefficient) distributions. The bootstrapping method generates N bootstrapping samples NPC and TPC by repeatedly sampling with replacement from the original NPC_{ij} and TPC_{ij} (e.g., Step 4), respectively. It can give us an empirical distribution of P-value θ, with which, we can estimate the probability that the distribution of two PC vectors are different. In our computational experiment, for a gene pair, if its value of $\Pr(\theta < 0.05)$ was larger than 0.8, we considered it as a pair of genes with the correlation relation significantly different between normal and cancer cells.

Our method can be described as follows.

Step 1. Compute n correlation coefficient Matrices NPC^1–NPC^n from the normal samples in n datasets for every gene pairs. For example, NPC^1 is an $m \times m$ Matrix from normal samples in the first dataset, and NPC_{ij}^1 represents the correlation coefficient between gene i, and gene j.

Step 2. Compute n correlation coefficient Matrices TPC^1–TPC^n from the tumor samples in the n datasets for every gene pair.

Step 3. For every gene pair (gene i and gene j), let

$$\begin{aligned} \mathrm{NPC}_{ij} &= \left[\mathrm{NPC}_{ij}^1 \ \ \mathrm{NPC}_{ij}^2 \ \ \mathrm{NPC}_{ij}^3 \ \ \cdots \ \ \mathrm{NPC}_{ij}^n\right], \\ \mathrm{TPC}_{ij} &= \left[\mathrm{TPC}_{ij}^1 \ \ \mathrm{TPC}_{ij}^2 \ \ \mathrm{TPC}_{ij}^3 \ \ \cdots \ \ \mathrm{TPC}_{ij}^n\right], \end{aligned} \tag{9}$$

Step 4. Perform the following (N is the number of samples we will generate using bootstrapping).

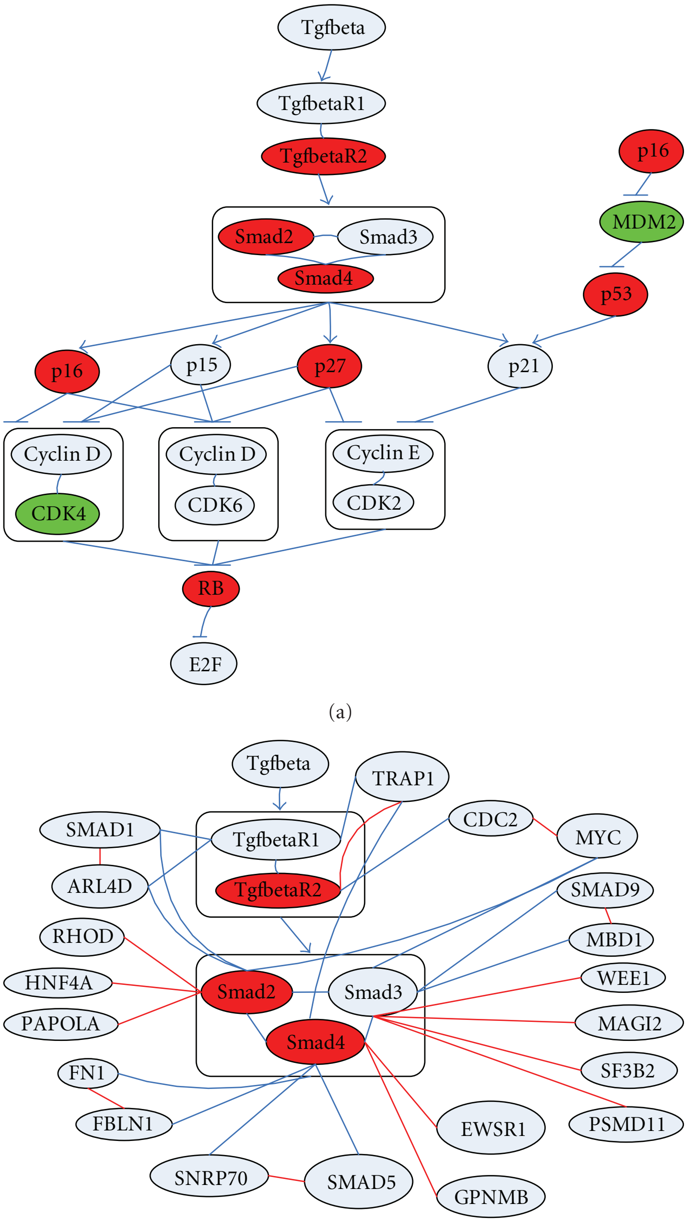
```
for $k = 1$ to $N$
Do generate bootstrap samples NPC and TPC from $\mathrm{NPC}_{ij}$ and $\mathrm{TPC}_{ij}$, respectively.
$\theta_k$ = $P$-value of K-S test on NPC and TPC.
End-for
Output $\Pr(\theta < 0.05) = \sharp\ (\theta < 0.05)/N$.
```

3. Experimental Results

In this section, we applied the bootstrapping K-S test method to analyze three cancer related pathways.

3.1. Antigrowth Signaling Pathway. Antigrowth signals can control proliferation in normal samples. Cancer cells have the ability to evade these antiproliferation signals. In the antigrowth signaling pathway, transforming growth factor beta (TGFβ) initiates this pathway by binding to two TGFβ receptors, Tgfbr1 and Tgfbr2. These two activated Tgfβ receptors can phosphorylate Smad2, Smad3, and Smad4 [28]. The SMAD family proteins then transduce antigrowth signals to the cell cycle inhibitors p21, p16, p27, and p15, which can inhibit the action of cyclin-CDK complex. The cyclin-CDK complex can phosphorylate RB and make RB dissociate from the E2F/RB complex to liberate E2F to activate the cell cycle procession from G1 to S phase (Figure 1(a)).

There are 19 genes in the antigrowth signaling pathway (Figure 1(a)). We found 689 unique genes related to these 19 genes from TRANSFAC and HPRD. Among these 708 genes, there were 4215 paired gene interactions, among which the correlation relations of 47 gene pairs were identified as significantly changed between normal and cancer cells. Among these 47 relations, we detected a cluster around SMAD family proteins which contained 15 relations with different distributions between normal samples and tumor samples (Figure 1(b)). Most of them came from large-scale protein-protein interaction experiments without the associated molecular function. For example, (Smad1–Arl4d), (RHOD–Smad2), and (WEE1–Smad3) in [29], (PAPOLA–Smad2), (SNRP70–Smad5), (GPNMB–Smad4), (PSMD11–Smad3), and (Smad9–MBD1) in [30], and (EWSR1–Smad4) in [31], all of them were detected based on large-scale protein-protein interaction experiments without annotation



(a)

(b)

Figure 1: Antigrowth signaling pathway and cluster around SMAD proteins. (a) Antigrowth signaling pathway. Nodes and edges represent human proteins and protein-protein interactions, respectively. Edges with direction represent a regulatory relation. → means an activating relation and, ⊣ means an inhibitory relation. (b) Cluster around smads. Red edges represent differentially changed relations. Blue edges represent unchanged relations. Red nodes represent tumor suppressor genes, and green nodes represent oncogenes.

of molecular function. Our results indicate that although their associated functions and internal mechanisms are still unclear, these gene pairs are related to the TGFβ-SMAD signaling pathway, and the relation between the two genes in each pair is significantly different in cancer and normal cells. Additionally, we identified some differentially changed relations with known molecular functions as follows:

(1) MAGI2 (a.k.a. ARIP1)–Smad3. MAGI2 (ARIP1) can interact with Smad3, and overexpression of ARIP1 can significantly suppress Smad3-induced transcriptional activity [32]. We validated this from the boxplot for MAGI2 (ARIP1)–Smad3 (Figure 2(a)). In normal samples, MAGI2 (ARIP1) and Smad3 showed a high positive correlation, while they had a high negative correlation in tumor samples.

(2) EWSR1–Smad4. Although the experiment type of the interaction between EWSR1 and Smad4 is yeast two-hybrid [31], mutations in EWSR1 are known to cause Ewing sarcoma and other members of the

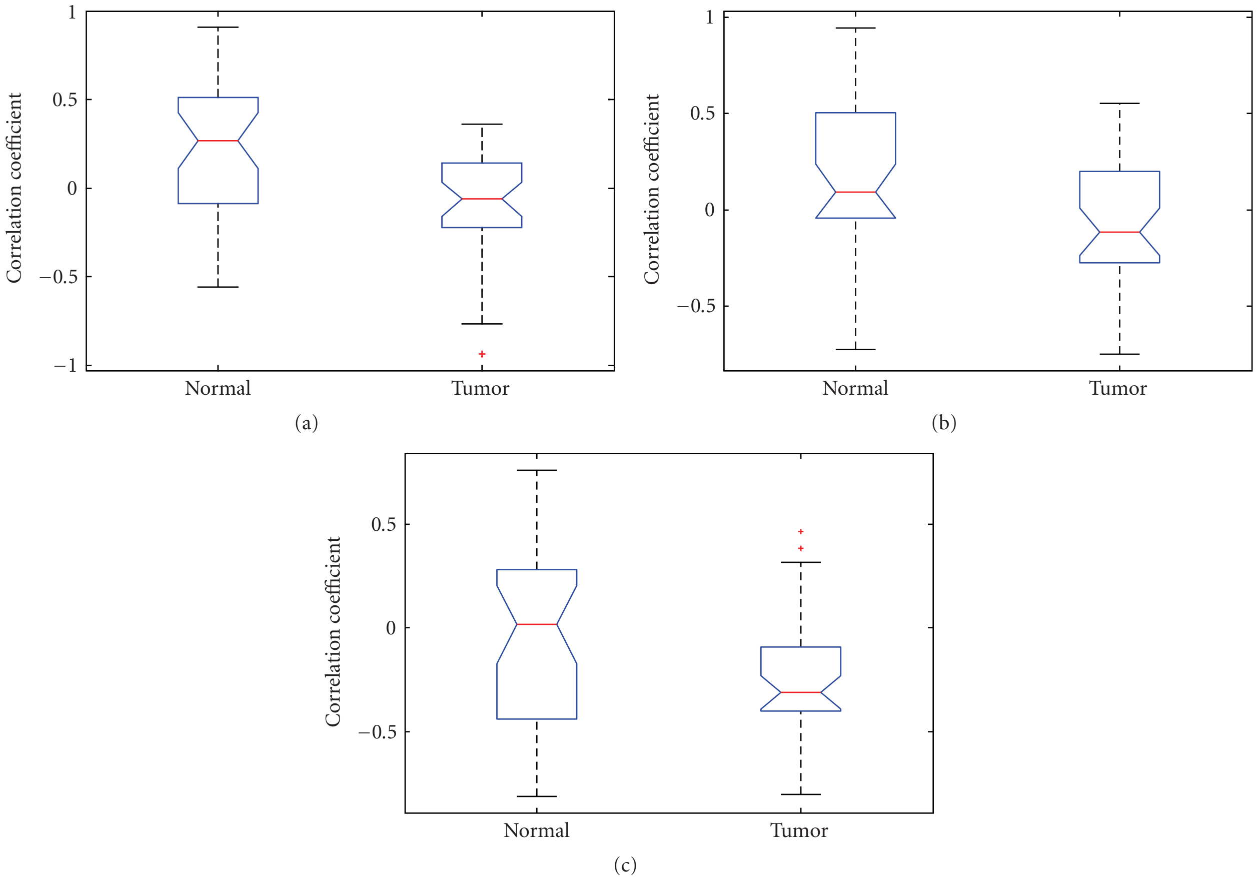


Figure 2: (a) Boxplot for MAGI2 (ARIP1)–Smad3. $\Pr(\theta < 0.05) = 0.986$. (b) Boxplot for EWSR1–Smad4. $\Pr(\theta < 0.05) = 0.954$. (c) Boxplot for TRAP1–TgfbetaR2. $\Pr(\theta < 0.05) = 0.944$.

Ewing family of tumors [33]. From the boxplot for EWSR1–Smad4, we found that the third quartile is the densest part of the whole distribution for both normal and tumor samples. The third quartile for normal samples showed a positive correlation whereas that for tumor samples showed a negative correlation (Figure 2(b)). Therefore, we suspect that EWSR1 can suppress the activity of Smad4 in tumor samples.

(3) TRAP1–Tgfbr2. TRAP1 has been shown to bind to TGFβ receptors and play a role in TGFβ signaling pathway. TRAP1 can interact with Smad4 and affect the SMAD-mediated signal transduction pathway. Mutant TRAP1 can prevent the formation of the Smad2–Smad4 complex to inhibit the TGFβ Signaling pathway [34]. In the boxplot for TRAP1–Tgfbr2 (Figure 2(c)), the densest quartile for tumor samples showed a high negative correlation.

3.2. Apoptosis Pathway. Cancer cells have the ability to evade programmed cell death or apoptosis. TNFα, FASL, TRAIL, and other genes can initiate apoptosis by binding to their receptors such as TNFR1, FAS, and TRAIL-R. Many apoptosis signals induce mitochondrial changes. Mitochondria can help transduce the apoptosis signals by releasing cytochrome C (Cytc), a potent catalyst of apoptosis. There are two different Bcl-2 family members: proapoptotic members (Bid, BAD) and antiapoptotic members (Bcl-2, Bcl-xl), which activate and inhibit, respectively, the release of Cytc. Finally, two key caspases (Casp8 and Casp9) activate other downstream caspases that perform the cascading events of cell death (Figure 3(a)).

In our results, we detected 33 relations with different distributions in the apoptosis pathway, and some are supported by existing experimental evidence. Examples include (Figure 3(b)) the following:

(1) PUMA–Bcl-XL (BCL2L1). PUMA can interact with Bcl-XL and meanwhile PUMA can also neutralize and antagonize all the Bcl-2-like proteins [35]. From the boxplot for PUMA–Bcl-XL, we can find that Bcl-XL, and PUMA showed a higher negative correlation in normal samples than in tumor samples (Figure 4(a)).

(2) AKT1–BAD. Active forms of Akt can phosphorylate BAD in vivo and in vitro to prevent it from promoting cell death [36]. In the boxplot for AKT1–BAD, the first quartile, the densest quartile for normal samples, showed a higher positive correlation than the second

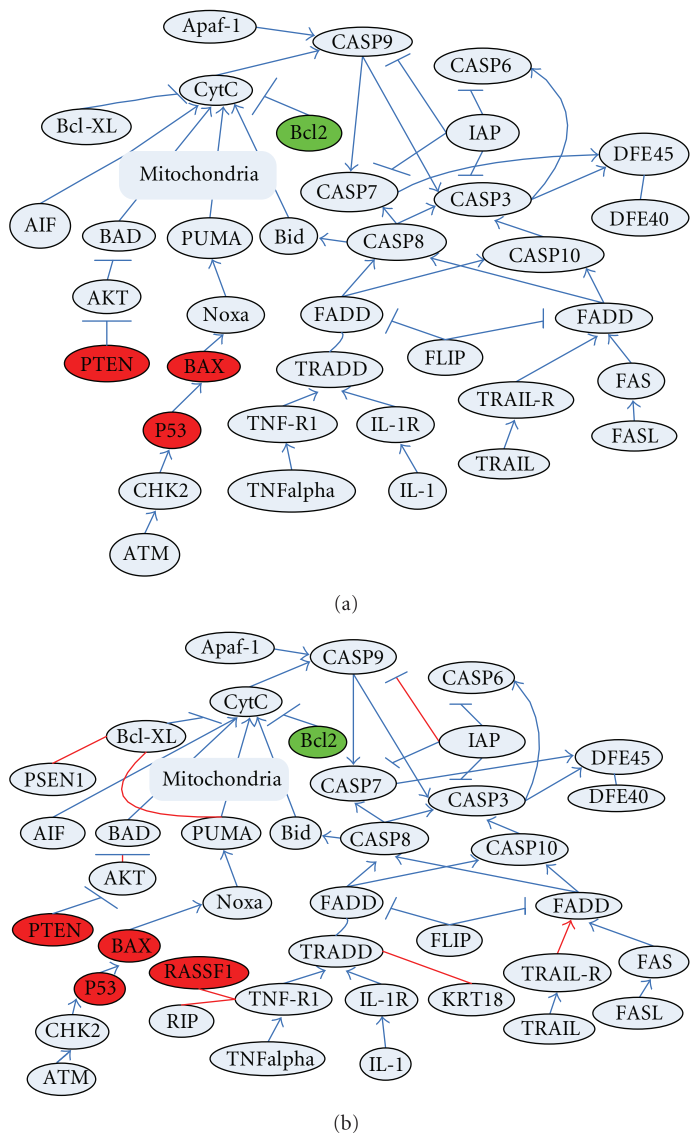


Figure 3: (a) Apoptosis pathway. (b) Differentially changed gene relations in apoptosis pathway. Red edges represent differentially changed relations. Blue edges represent unchanged relations. Red nodes represent tumor suppressor genes, and green nodes represent oncogenes.

quartile, the densest for tumor samples (Figure 4(b)). So we speculated that Akt can suppress BAD's activity in tumor samples.

(3) KRT18–TRADD. TRADD is a KRT18-interacting protein. KRT18 may inactivate TRADD to prevent interactions between TRADD and the activated TNFR1 and thus affect TNFα-induced apoptosis [37]. In the boxplot for KRT18–TRADD, normal samples showed a higher positive correlation (Figure 4(c)).

(4) TNFR1–RIPK1 (RIP). The interaction between the death domain of TNFα receptor-1 (TNFR1) and TRADD can trigger distinct signaling pathways leading to apoptosis. TRADD also interacts strongly with another death domain protein; RIP and RIP plays an important role in the TNF signaling cascades leading to apoptosis [38]. In the boxplot for TNFR1–RIPK1, TNFR1 and RIPK1 exhibited high positive correlation in normal samples (Figure 4(d)).

(5) TNFR1–RASSF1. RASSF1A is a tumor suppressor gene. Apoptosis initiation by TNFα or TRAIL recruits RASSF1A and MAP-1 to form complexes. RASSF1A and MAP-1 are the key links between death receptors and the apoptotic machinery [39]. This was verified by the Boxplot for TNFR1–RASSF1. In most normal samples, these genes showed a high positive correlation. In most tumor samples, they showed a zero or negative correlation (Figure 4(e)).

(6) IAP–CASP9. Inhibitor of apoptosis (IAP) suppresses the activities of caspases and inhibits different apoptotic pathways [40]. IAP and CASP9 showed a high negative correlation in tumor samples (Figure 4(f)).

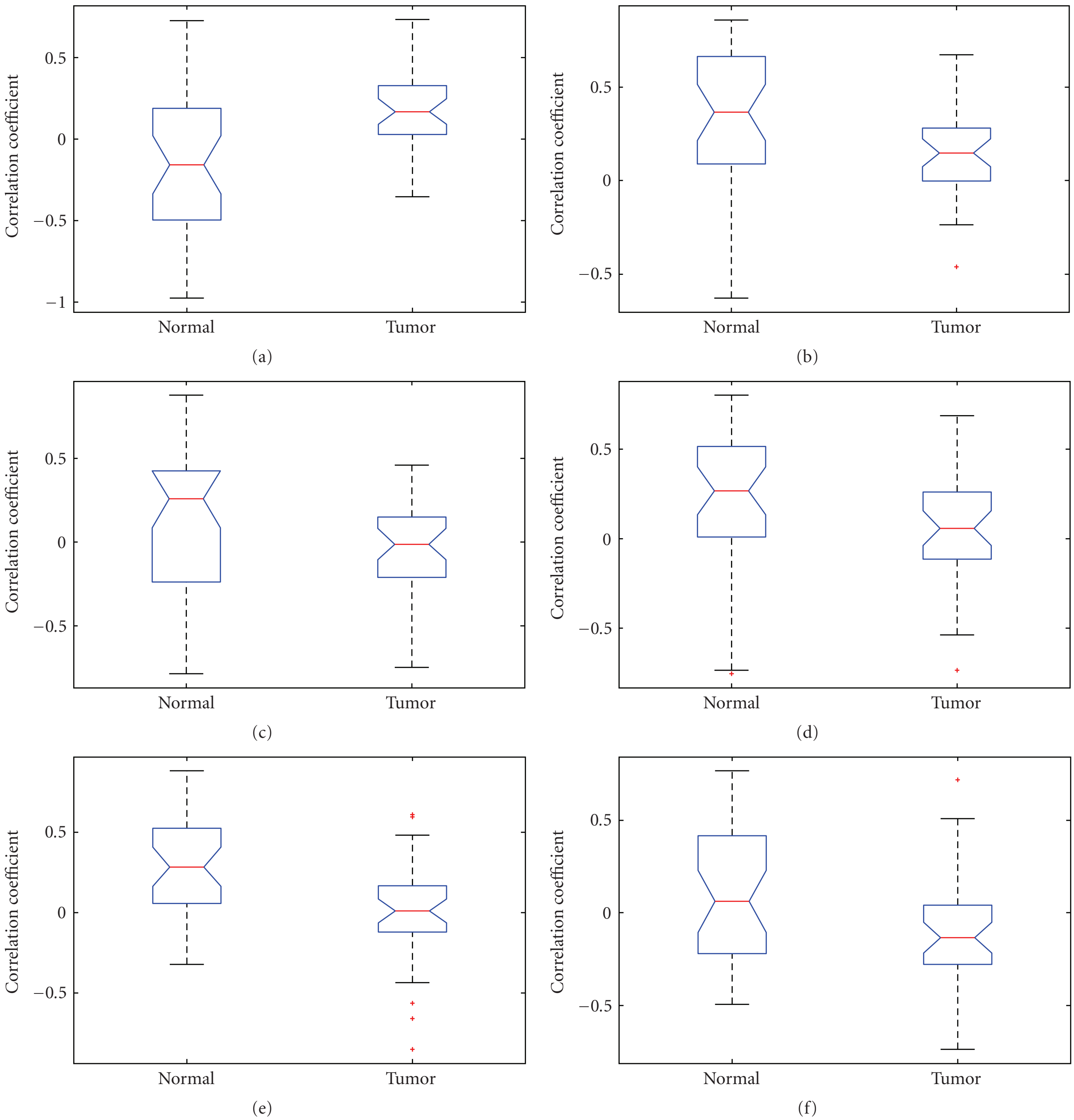


FIGURE 4: (a) Boxplot for PUMA–Bcl-XL(BCL2L1). $\Pr(\theta < 0.05) = 0.998$. (b) Boxplot for AKT1–BAD. $\Pr(\theta < 0.05) = 0.859$. (c) Boxplot for KRT18–TRADD. $\Pr(\theta < 0.05) = 0.991$.(d) Boxplot for TNFR1–RIPK1(RIP). $\Pr(\theta < 0.05) = 0.831$. (e) Boxplot for TNFR1–RASSF1. $\Pr(\theta < 0.05) = 0.946$. (f) Boxplot for IAP–CASP9. $\Pr(\theta < 0.05) = 0.826$.

Among the eight differential gene relations in Figure 3(b), three of them were in the seed pathway: TRAIL-R → FADD, IAP → CASP9, and AKT → BAD, which demonstrates the effectiveness of the proposed method.

3.3. Growth Signaling Pathway. Cancer cells have the ability to produce their own growth promoting signals. EGF, TGFα, and PDGF are activated and then bind to their receptors to transduce the growth signals. The activated growth factor receptors can in turn activate the SOS-Ras_Raf_Mapk cascade. In the growth signal pathway (Figure 5), Ras, JUN, and Fos are oncogenes.

We could find 68 relations with different distributions in the growth signal pathway, and we discuss three relations as follows:

(1) RASSF2–KRAS. Although different forms of Ras are frequently thought of as oncogenes, they also have the ability to produce antigrowth effects such as cell

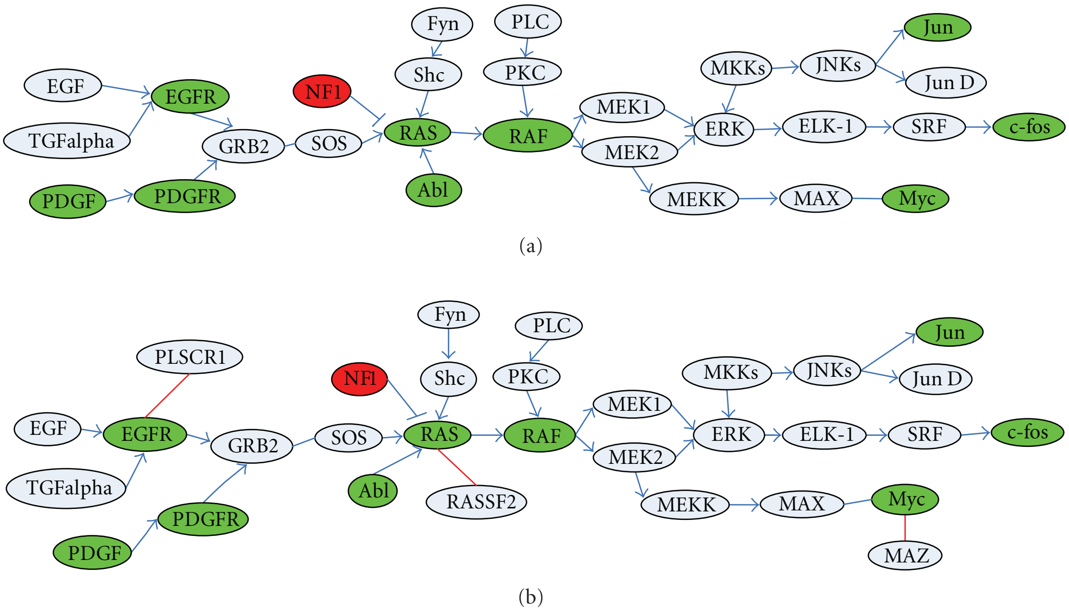


FIGURE 5: (a) Growth signal pathway. (b) Differentially changed relations in growth signal pathway. Red edges represent differentially changed relations. Blue edges represent unchanged relations. Red nodes represent tumor suppressor genes, and green nodes represent oncogenes.

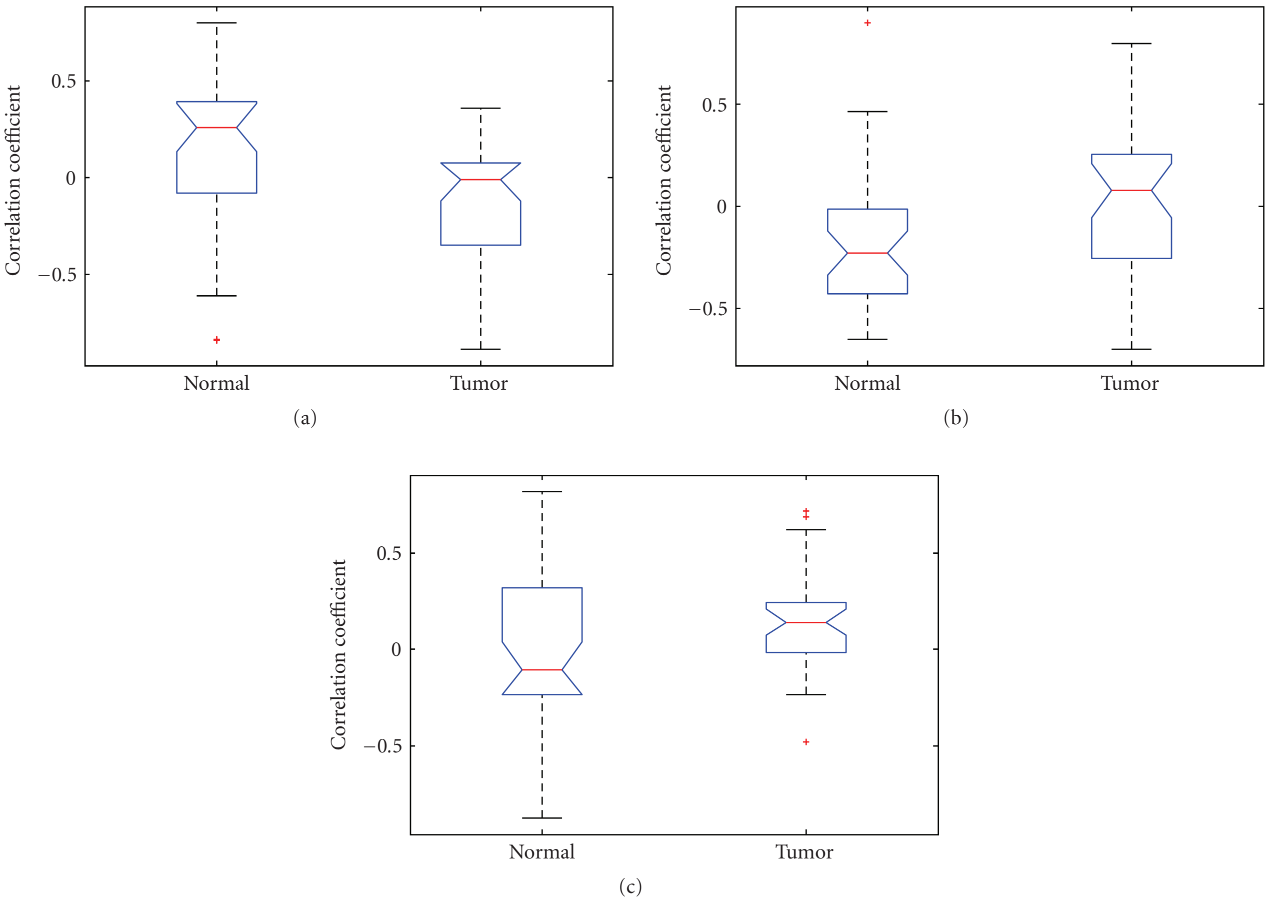


FIGURE 6: (a) Boxplot for RASSF2–KRAS. $\Pr(\theta < 0.05) = 0.983$. (b) Boxplot for MAZ–MYC. $\Pr(\theta < 0.05) = 0.833$. (c) Boxplot for PLSCR1–EGFR. $\Pr(\theta < 0.05) = 0.963$.

cycle arrest, differentiation, and apoptosis. RASSF2 can bind directly to K-Ras. Moreover, RASSF2 can inhibit the growth of tumor cells, and the activated K-Ras can enhance this ability [41]. This might be why RASSF2 and RAS showed a high positive correlation in normal samples in the boxplot (Figure 6(a)).

(2) MAZ–MYC. The MAZ family can increase the oncogene MYC's transcriptional activity [42]. As expected, MAZ and MYC demonstrated a higher positive correlation in tumor samples (Figure 6(b)).

(3) PLSCR1–EGFR. Activated epidermal growth factor receptors (EGFRs) can both physically and functionally interact with PLSCR1. In turn, PLSCR1 can interact with Shc and thus accelerate the activation of Src kinase through the EGF receptor, while Src can initiate some activating pathway for the oncogene JUN [43]. In the boxplot for PLSCR1–EGFR, the densest quartile for normal samples showed a low negative correlation, whereas the densest quartile for tumor samples showed a low positive correlation (Figure 6(c)).

4. Conclusion and Discussion

After several decades of cancer research, some details of the underlying mechanisms of cancer at the gene level are still unclear. In this paper, we propose an integrative method based on the bootstrapping K-S test to evaluate a large number of microarray datasets generated from 21 different types of cancer in order to identify gene pairs that have different relationships in normal versus cancer tissues. The significant alteration of gene relations can greatly extend our understanding of the molecular mechanisms of human cancer. In our method, we obviate the disadvantage of the traditional t-test, which only considers the mean and variance of samples and fails in the analysis of microarray data with small numbers of samples. Instead of the t-test, we propose the use of the bootstrapping K-S test method to detect gene pairs with different distributions of Pearson correlation coefficient values in normal and tumor samples. The experimental results demonstrated that our method could find meaningful alterations in gene relations and opened a potential door for further cancer research.

Acknowledgment

This work was supported by NSF award IIS-0644366.

References

[1] H. Han, D. J. Bearss, L. W. Browne, R. Calaluce, R. B. Nagle, and D. D. Von Hoff, "Identification of differentially expressed genes in pancreatic cancer cells using cDNA microarray," *Cancer Research*, vol. 62, no. 10, pp. 2890–2896, 2002.

[2] X.-W. Chen, "Margin-based wrapper methods for gene identification using microarray," *Neurocomputing*, vol. 69, no. 16–18, pp. 2236–2243, 2006.

[3] A. A. Margolin, I. Nemenman, K. Basso, et al., "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, supplement 1, article S7, pp. 1–15, 2006.

[4] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[5] X.-W. Chen, G. Anantha, and X. Wang, "An effective structure learning method for constructing gene networks," *Bioinformatics*, vol. 22, no. 11, pp. 1367–1374, 2006.

[6] H. Xiong and X.-W. Chen, "Kernel-based distance metric learning for microarray data classification," *BMC Bioinformatics*, vol. 7, article 299, pp. 1–11, 2006.

[7] T. Golub, D. Slonim, P. Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[8] H. Xiong, Y. Zhang, and X.-W. Chen, "Data-dependent kernel machines for microarray data classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 4, pp. 583–595, 2007.

[9] T.-M. Chu, B. Weir, and R. Wolfinger, "A systematic statistical linear modeling approach to oligonucleotide array experiments," *Mathematical Biosciences*, vol. 176, no. 1, pp. 35–51, 2002.

[10] W.-P. Hsieh, T.-M. Chu, R. D. Wolfinger, and G. Gibson, "Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles," *Genetics*, vol. 165, no. 2, pp. 747–757, 2003.

[11] M. Neuhäuser and R. Senske, "The Baumgartner-Weiß-Schindler test for the detection of differentially expressed genes in replicated microarrays experiments," *Bioinformatics*, vol. 20, no. 18, pp. 3553–3564, 2004.

[12] O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman, "Nonparametric methods for identifying differentially expressed genes in microarray data," *Bioinformatics*, vol. 18, no. 11, pp. 1454–1461, 2002.

[13] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[14] H. Yang and G. Churchill, "Estimating p-values in small microarray experiments," *Bioinformatics*, vol. 23, no. 1, pp. 38–43, 2007.

[15] Y. Zhao and W. Pan, "Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 19, no. 9, pp. 1046–1054, 2003.

[16] S. Draghici, P. Khatri, A. L. Tarca, et al., "A systems biology approach for pathway level analysis," *Genome Research*, vol. 17, no. 10, pp. 1537–1545, 2007.

[17] T. Manoli, N. Gretz, H.-J. Gröne, M. Kenzelmann, R. Eils, and B. Brors, "Group testing for pathway analysis improves comparability of different microarray datasets," *Bioinformatics*, vol. 22, no. 20, pp. 2500–2506, 2006.

[18] K.-C. Li, "Genome-wide coexpression dynamics: theory and application," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 26, pp. 16875–16880, 2002.

[19] Y. Lai, B. Wu, L. Chen, and H. Zhao, "A statistical method for identifying differential gene-gene co-expression patterns," *Bioinformatics*, vol. 20, no. 17, pp. 3146–3155, 2004.

[20] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: is there a unique set?" *Bioinformatics*, vol. 21, no. 2, pp. 171–178, 2005.

[21] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature Reviews Genetics*, vol. 7, no. 1, pp. 55–65, 2006.

[22] J. K. Choi, U. Yu, O. J. Yoo, and S. Kim, "Differential coexpression analysis using microarray data and its application to human cancer," *Bioinformatics*, vol. 21, no. 24, pp. 4348–4355, 2005.

[23] T. Barrett, T. O. Suzek, D. B. Troup, et al., "NCBI GEO: mining millions of expression profiles—database and tools," *Nucleic Acids Research*, vol. 33, database issue, pp. D562–D566, 2005.

[24] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, 2000.

[25] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.

[26] V. Matys, E. Fricke, R. Geffers, et al., "TRANSFAC®: transcriptional regulation, from patterns to profiles," *Nucleic Acids Research*, vol. 31, no. 1, pp. 374–378, 2003.

[27] W. J. Conover, *Practical Nonparametric Statistics*, John Wiley & Sons, New York, NY, USA, 3rd edition, 1999.

[28] G. C. Blobe, X. Liu, S. J. Fang, T. How, and H. F. Lodish, "A novel mechanism for regulating transforming growth factor β (TGF-β) signaling: functional modulation of type III TGF-β receptor expression through interaction with the PDZ domain protein, GIPC," *The Journal of Biological Chemistry*, vol. 276, no. 43, pp. 39608–39617, 2001.

[29] M. Barrios-Rodiles, K. R. Brown, B. Ozdamar, et al., "High-throughput mapping of a dynamic signaling network in mammalian cells," *Science*, vol. 307, no. 5715, pp. 1621–1625, 2005.

[30] F. Colland, X. Jacq, V. Trouplin, et al., "Functional proteomics mapping of a human signaling pathway," *Genome Research*, vol. 14, no. 7, pp. 1324–1332, 2004.

[31] J.-F. Rual, K. Venkatesan, T. Hao, et al., "Towards a proteome-scale map of the human protein-protein interaction network," *Nature*, vol. 437, no. 7062, pp. 1173–1178, 2005.

[32] H. Shoji, K. Tsuchida, H. Kishi, et al., "Identification and characterization of a PDZ protein that interacts with activin type II receptors," *The Journal of Biological Chemistry*, vol. 275, no. 8, pp. 5485–5492, 2000.

[33] J. Ban, C. Siligan, M. Kreppel, D. Aryee, and H. Kovar, "EWS-FLI1 in Ewing's sarcoma: real targets and collateral damage," *Advances in Experimental Medicine and Biology*, vol. 587, pp. 41–52, 2006.

[34] J. U. Wurthner, D. B. Frank, A. Felici, et al., "Transforming growth factor-β receptor-associated protein 1 is a Smad4 chaperone," *The Journal of Biological Chemistry*, vol. 276, no. 22, pp. 19495–19502, 2001.

[35] L. Chen, S. N. Willis, A. Wei, et al., "Differential targeting of prosurvival Bcl-2 proteins by their BH3-only ligands allows complementary apoptotic function," *Molecular Cell*, vol. 17, no. 3, pp. 393–403, 2005.

[36] L. del Peso, M. González-García, C. Page, R. Herrera, and G. Nuñez, "Interleukin-3-induced phosphorylation of BAD through the protein kinase Akt," *Science*, vol. 278, no. 5338, pp. 687–689, 1997.

[37] H. Inada, I. Izawa, M. Nishizawa, et al., "Keratin attenuates tumor necrosis factor-induced cytotoxicity through association with TRADD," *The Journal of Cell Biology*, vol. 155, no. 4, pp. 415–426, 2001.

[38] H. Hsu, J. Huang, H.-B. Shu, V. Baichwal, and D. V. Goeddel, "TNF-dependent recruitment of the protein kinase RIP to the TNF receptor-1 signaling complex," *Immunity*, vol. 4, no. 4, pp. 387–396, 1996.

[39] S. Baksh, S. Tommasi, S. Fenton, et al., "The tumor suppressor RASSF1A and MAP-1 link death receptor signaling to bax conformational change and cell death," *Molecular Cell*, vol. 18, no. 6, pp. 637–650, 2005.

[40] Q. L. Deveraux, N. Roy, H. R. Stennicke, et al., "IAPs block apoptotic events induced by caspase-8 and cytochrome c by direct inhibition of distinct caspases," *The EMBO Journal*, vol. 17, no. 8, pp. 2215–2223, 1998.

[41] M. D. Vos, C. A. Ellis, C. Elam, A. S. Ülkü, B. J. Taylor, and G. J. Clark, "RASSF2 is a novel K-Ras-specific effector and potential tumor suppressor," *The Journal of Biological Chemistry*, vol. 278, no. 30, pp. 28045–28051, 2003.

[42] H. Tsutsui, O. Sakatsume, K. Itakura, and K. K. Yokoyama, "Members of the MAZ family: a novel cDNA clone for MAZ from human pancreatic islet cells," *Biochemical and Biophysical Research Communications*, vol. 226, no. 3, pp. 801–809, 1996.

[43] M. Nanjundan, J. Sun, J. Zhao, Q. Zhou, P. J. Sims, and T. Wiedmer, "Plasma membrane phospholipid scramblase 1 promotes EGF-dependent activation of c-Src through the epidermal growth factor receptor," *The Journal of Biological Chemistry*, vol. 278, no. 39, pp. 37413–37418, 2003.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 892863, 14 pages
doi:10.1155/2009/892863

Research Article

Bioinformatics Methods for Learning Radiation-Induced Lung Inflammation from Heterogeneous Retrospective and Prospective Data

Sarah J. Spencer,[1] Damian Almiron Bonnin,[2] Joseph O. Deasy,[1] Jeffrey D. Bradley,[1] and Issam El Naqa[1]

[1] Department of Radiation Oncology, Washington University Medical School, Saint Louis, MO 63110, USA
[2] Biochemistry Department, Earlham College, Richmond, IN 47374, USA

Correspondence should be addressed to Issam El Naqa, elnaqa@wustl.edu

Received 14 January 2009; Accepted 10 March 2009

Recommended by Zhenqiu Liu

Radiotherapy outcomes are determined by complex interactions between physical and biological factors, reflecting both treatment conditions and underlying genetics. Recent advances in radiotherapy and biotechnology provide new opportunities and challenges for predicting radiation-induced toxicities, particularly radiation pneumonitis (RP), in lung cancer patients. In this work, we utilize datamining methods based on machine learning to build a predictive model of lung injury by retrospective analysis of treatment planning archives. In addition, biomarkers for this model are extracted from a prospective clinical trial that collects blood serum samples at multiple time points. We utilize a 3-way proteomics methodology to screen for differentially expressed proteins that are related to RP. Our preliminary results demonstrate that kernel methods can capture nonlinear dose-volume interactions, but fail to address missing biological factors. Our proteomics strategy yielded promising protein candidates, but their role in RP as well as their interactions with dose-volume metrics remain to be determined.

1. Introduction

Lung cancer is one of the most lethal diseases among men and women worldwide. Patients suffering from lung cancer display a 5-year survival rate of only 15%, a value that has held constant over the past 30 years. According to the American Cancer Society (ACS) statistics, 215.020 new lung cancer cases and 161.840 deaths due to lung cancer are expected in the year 2008 alone [1]. This accounts for 29% of all cancer deaths with 87% of these cases classified clinically as nonsmall cell lung cancer (NSCLC). A large percentage of lung cancer patients receive radiation therapy (radiotherapy) as part of their standard of care and it is the main treatment for inoperable patients at advanced stages of the disease. Radiotherapy is a directed and localized treatment, but its dose is limited by toxicities to surrounding normal tissues. Thus, patients are at risk of experiencing tumor recurrence if insufficient dose was prescribed or conversely they are susceptible to toxicities if exposed to excessive doses.

The last two decades have witnessed many technological advances in the development of three-dimensional treatment planning systems and image-guided methods to improve tumor localization while sparing surrounding normal tissues [2, 3]. In parallel, there has been a tremendous evolution in biotechnology providing high-throughput genomics and proteomics information applicable within cancer radiation biology. This has led to the birth of a new field in radiation oncology denoted as "radiogenomics" or "radioproteomics" [4, 5]. These advances, if directed properly, could pave the way for increasingly individualized and patient-specific treatment planning decisions that continue to draw from estimates of tumor local control probability (TCP) or surrounding normal tissues complication probability (NTCP) as illustrated in Figure 1.

Traditionally, tissue radioresponse has been modeled using simplistic expressions of cell kill based on the linear-quadratic (LQ) model developed in the 1940s [6]. The LQ

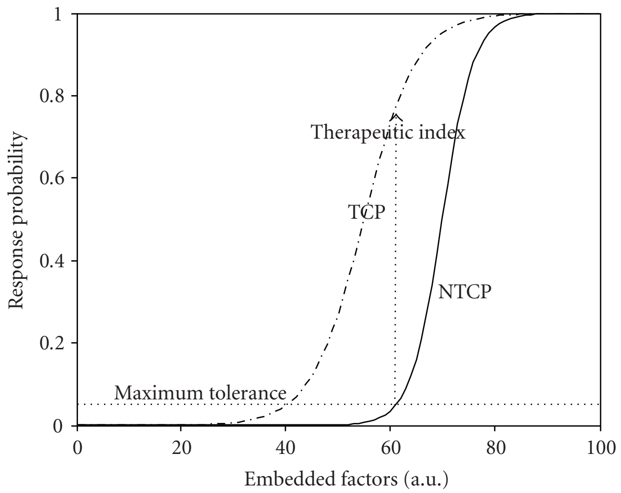


Figure 1: An S-shaped response curves representing tumor control probability (TCP) and normal tissue complication probability (NTCP) postradiotherapy as a function of treatment factors. The probabilities could be constructed as a function of heterogeneous variables (dose-volume metrics, biomarkers, and clinical factors). The radiotherapy treatment objective is to maximize the therapeutic index for each patient case.

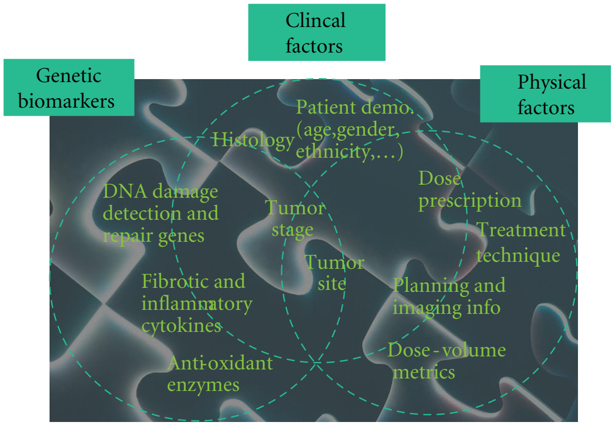


Figure 2: Radiotherapy treatment involves complex interaction of physical, biological, and clinical factors. The successful datamining approach should be able to resolve this interaction "puzzle" in the observed treatment outcome (e.g., radiation-induced lung injury) for each individual patient.

formalism describes repairable and nonrepairable radiation damage of different tissue types with a few estimated radiation sensitivity parameters from cell culture assays [7]. Despite the historical value of LQ-based models, several authors have recently cautioned against its limitations [8, 9]. It is understood that radiotherapy outcomes are determined by complex interactions between physical treatment factors, anatomical structures, and patient-related genetic variables as depicted in Figure 2.

A different approach based on datamining of patient information (clinical, physical, and biological records) has been proposed to ameliorate these challenges and bridge the gap between traditional radiobiological predictions from in vitro assays and observed treatment outcomes in clinical practice by understanding the underlying molecular mechanisms [10–12]. The main idea of data-driven models is to utilize datamining approaches and statistical model building methods to integrate disparate predictive factors. Such models may improve predictive power, but they must be simultaneously guarded for overfitting pitfalls using resampling techniques, for instance. This approach is motivated by the extraordinary increase in patient-specific biological and clinical information from progress in genetics and imaging technology. The main goal is to resolve the complicated interactions by proper mixing of heterogeneous variables (Figure 2). As a result, the treatment planning system could be optimized to yield the best possible care for the patient as illustrated in Figure 3.

Most data-driven models in the radiation oncology literature could be categorized into two types of models: (1) physical dose-volume models or (2) single-biomarkers models. Dose-volume models are driven by the presence of large treatment planning archives and the current clinical practice of radiotherapy treatment. Current radiotherapy protocols allow for the extraction of parameters that relate irradiation dose to the treated volume fractions (tumors or surrounding normal organs at risk) in dose-volume histograms [13]. Conversely, screening for different blood/tissue biomarkers to predict radiation response (TCP or NTCP) is an emerging field in radiation oncology with many promising opportunities as well as new technical challenges regarding data collection quality, the advancement of lab techniques, and the development of statistical methodology [14].

To illustrate and investigate the changing landscape of radiation response modeling, our study addresses radiation pneumonitis (RP), the major dose limiting toxicity in thoracic irradiation. Clinically, RP is lung inflammation that usually occurs within six months after therapy for a subset of patients and can manifest as cough, dyspnea, fever, and/or malaise which may require significant supportive measures including steroids and oxygen supplementation [15]. In its worst form, RP can continue to progress and result in death. According to the NCI Common Terminology Criteria for Adverse Events (CTCAEs) v3.0, a clinical scoring system for RP, the severity of pneumonitis is graded from 0 (minimal symptoms) to 4 (most severe/life-threatening) or even 5 (death). A CTCAE-v3.0 grade ≥3 indicates clinical onset of severe RP. Biologically, the ionizing radiation from treatment can cause damage to the normal alveolar epithelium cells (airways) of the lung resulting in release of a wetting agent surfactant into the alveolar space and detachment of the pneumocytes from their basement membrane. It is thought that this process triggers a cascade of humoral cellular and immune response events among alveolar epithelium, fibroblasts, lymphocytes, and macrophages leading to RP as shown in Figure 4 [16].

We conjecture that a good predictive model for radiation hypersensitivity should be able to properly describe the interactions between physical and biological processes resulting from radiation exposure and adequately span the variable

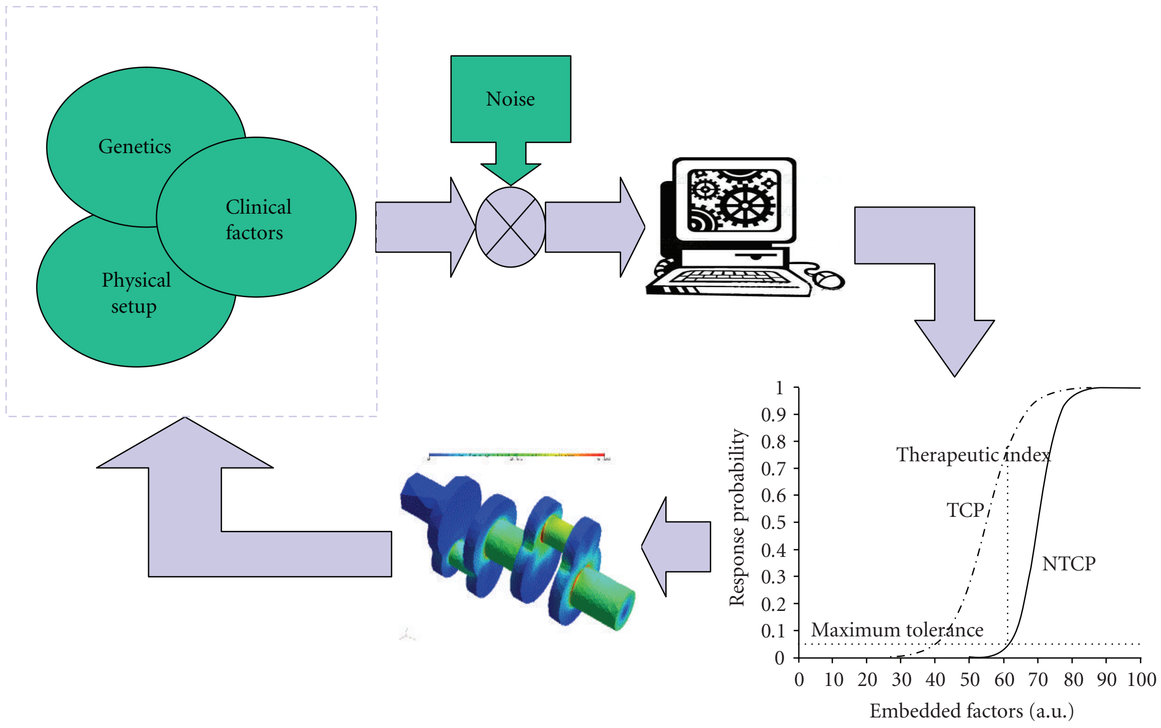


Figure 3: The datamining understanding of these heterogeneous variables interactions could be fed back into the treatment planning system to improve patient's outcomes.

space shown in Figure 2. Working towards this standard, we will present our utilization of supervised and unsupervised machine learning approaches to interrogate radiation oncology data and develop methodology for building better predictive models of radiation therapy response. We start by examining existing treatment planning archives and conduct retrospective analysis of physical dose-volume models to predict the onset of RP. We then describe our attempt to fillin the prediction gap in such physical models through a prospective study that considers preexisting biological variables, which may influence treatment response. Note that the retrospective study has the advantage of large sample size and hence higher power while the prospective approach is focused towards improving current prediction by incorporating missing information in past archives into more comprehensive databases and performing evaluation on new unseen data. In particular, we will present our proteomic methodology to investigate predictive biomarkers of RP that could eliminate informational gaps in our retrospective physical model.

The paper is organized as follows. In Section 2, we describe our retrospective analysis of dose-volume RP predictors and our current prospective proteomic analysis. In Section 3, we contrast our results using model-building approaches based on logistic regression, support vector machine, and a 3-way design for biomarker discovery in proteomic analysis of RP. Methods for variable selections are analyzed. Lastly, in Section 4 we discuss our current findings and offer some concluding remarks in Section 5.

2. Materials and Methods

2.1. Dataset Description. To demonstrate our methodology, separate datasets were compiled using data from two groups of patients all diagnosed with nonsmall cell lung cancer (NSCLC) and treated with three-dimensional conformal radiation therapy (3D-CRT) at our institution. The first dataset was collected retrospectively from the clinical archives with median doses around 70 Gy (the doses were corrected to account for lung heterogeneity using the tissue-air ratio method). In this set, 52 out of 219 patients were diagnosed with postradiation late pneumonitis (RTOG grade ≥3). The dataset included clinical and dosimetric (dose-volume) variables. The clinical variables included age, gender, ethnicity, date of treatment start, treatment technique, treatment aim, chemotherapy, disease stage, treatment duration, histological features, and so forth. The dosimetric variables compiled for this retrospective dataset were measured and calculated in reference to the extensive dose-volume documentation in the radiation oncology literature. Typically, these metrics are extracted from the dose-volume histogram (DVH) and include V_x (the percentage volume that got x Gy), D_x (the minimum dose to the hottest x% volume), mean dose, maximum and minimum doses, generalized equivalent uniform, and so forth. In-house software tools for data dearchiving, the analysis software a Computational Environment for Radiotherapy Research (CERR) [17], and the dose response explorer system (DREES) [18] were used to extract the different metrics and analyze their association with RP.

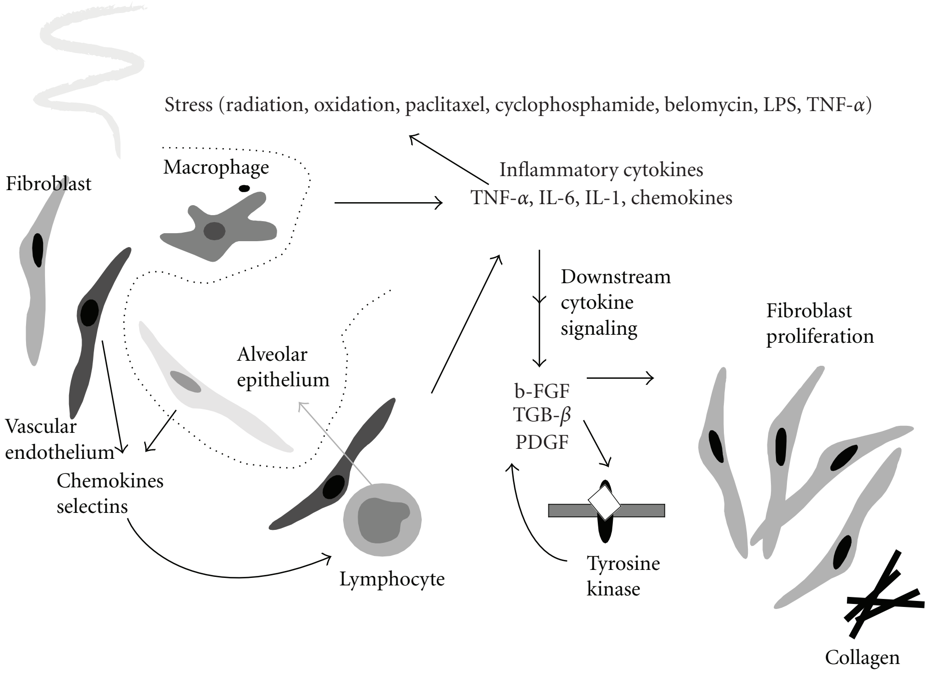


Figure 4: A schematic diagram of the possible cellular and molecular events involved in pulmonary injury by radiation. Cellular interaction among endothelial cells, alveolar epithelial cells, macrophages, lymphocytes, and fibroblasts, through cytokine mediators of chemokines, selectins, inflammatory cytokines, and fibrotic cytokines are involved (from Chen et al., Seminars in Surgical Oncology, 2003).

The second dataset was collected from September 2007 to September 2008 for a prospective analysis. Nineteen patients were involved in the study and underwent conventional radiotherapy with mean doses close to 70 Gy. Out of nineteen patients, four were diagnosed with postradiation late pneumonitis (RTOG grade ≥3). The data collected for each patient included the same clinical and dosimetric variables as the prospective study. In addition to this data, five blood samples were drawn from each patient over the course of treatment. These sample collections were scheduled before radiotherapy (pretreatment), midtreatment, immediately after radiotherapy (posttreatment), and also at a three month and at six-month follow-up appointments.

This second dataset is gathered from an institutionally approved prospective study for extracting biomarkers to predict radiotherapy response in inoperable stage III NSCLC patients who receive radiotherapy as part of their treatment. For our preliminary proteomic screening, we selected two lung cancer patients who were treated using fractionated radiotherapy according to our institute clinical standards. One case was designated as *control* and the other case was for a patient who developed RP and designated as *disease.* The control patient, despite radiation treatment for advanced lung cancer, developed no adverse health conditions throughout a follow-up period of 14 months. RP typically occurs within the first year posttreatment with a mode of 6 months. The disease case selected for the study died due to a severe RP episode one month after the end of treatment. For both the control and disease cases, a serum sample drawn before treatment as well as a sample drawn at the last available follow-up was submitted for liquid chromatography mass spectrometry (LC-MS) analysis. A Seppro 15 × 13 mm chromatography column (LC20) (GenWay Biotech Inc., San Diego, Calif, USA) was used to deplete the thawed samples of the 14 most abundant proteins in human blood serum. The samples then underwent digestion by the serine protease trypsin with a 10 μg Bovine Serum Albumin (BSA) external standard. Subsequent LC-MS allowed for the separation and mass analysis of tryptic peptides in each of the four samples. The most abundant peptides of each MS mass scan were automatically sent to a second mass spectrometer for fragmentation and sequence determination according to a tandem MS (MS/MS) design.

2.2. Model Building Approach. In the context of data-driven outcomes modeling, the observed treatment outcome (e.g., normal tissue complication probability (NTCP) or tumor control probability (TCP)) is considered as the result of functional mapping of multiple dosimetric, clinical, or biological input variables [19]. Mathematically, this could be expressed as $f(\mathbf{x};\mathbf{w}^*) : X \rightarrow Y$, where $x_i \in \mathbb{R}^d$ are the input explanatory variables (dose-volume metrics, patient disease specific prognostic factors, or biological markers) of length d, $y_i \in Y$ are the corresponding observed treatment outcome (TCP or NTCP), and $\mathbf{w}^*$ includes the optimal parameters of outcome model $f(\cdot)$ obtained by optimizing a certain objective criteria. In our previous work [10, 19], a logit transformation was used as follows:

$$f(\mathbf{x}_i) = \frac{e^{g(\mathbf{x}_i)}}{1 + e^{g(\mathbf{x}_i)}}, \quad i = 1, \dots, n, \tag{1}$$

where n is the number of cases (patients), $\mathbf{x}_i$ is a vector of the input variable values used to predict $f(\mathbf{x}_i)$ for outcome y_i of the ith patient. The "x-axis" summation $g(\mathbf{x}_i)$ is given by

$$g(\mathbf{x}_i) = \beta_o + \sum_{j=1}^{d} \beta_j x_{ij}, \quad i = 1, \ldots, n, \; j = 1, \ldots, d, \tag{2}$$

where d is the number of model variables and the β's are the set of model coefficients determined by maximizing the probability that the data gave rise to the observations. A major weakness in using this formulation, however, is that the model capacity to learn is limited. In addition, (2) requires the user feedback to determine whether interaction terms or higher order terms should be added, making it a trial and error process. A solution to ameliorate this problem is offered by applying machine learning methods as discussed in the next section**.**

2.3. Kernel-Based Methods. Kernel-based methods and their most prominent member, support vector machines (SVMs), are universal constructive learning procedures based on the statistical learning theory [20]. These methods have been applied successfully in many diverse areas [21–25].

Statistical Learning. Learning is defined in this context as estimating dependencies from data [26]. There are two common types of learning: supervised and unsupervised. Supervised learning is used to estimate an unknown (input, output) mapping from known (input, output) samples (e.g., classification or regression). In unsupervised learning, only input samples are given to the learning system (e.g., clustering or dimensionality reduction). In this study, we focus mainly on supervised learning, wherein the endpoints of the treatments such as tumor control or toxicity grade are provided by experienced oncologists following RTOG or NCI criteria. Nevertheless, we will use unsupervised methods such as principle component analysis and multidimensional scaling to aid visualization of multivariate data and guide the selection of proper schemes for data analysis.

The main objective of supervised learning is to estimate a parametric function $f(\mathbf{x};\mathbf{w}^*) : X \rightarrow Y$ by assistance from a representative training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$. The two main supervised learning tasks are classification and regression. The difference between classification and regression is that the output y in case of classification belongs to a discrete, or categorical, set $y \in \{1, 2, \ldots, M\}$ (e.g., in binary classification $M = 2$), whereas in regression y is a continuous variable. In the example of classification (i.e., discrimination between patients who are at low risk versus patients who are at high risk of radiation pneumonitis), the main function of the kernel-based technique would be to separate these two classes with "hyperplanes" that maximize the margin (separation) between the classes in the nonlinear feature space defined by implicit kernel mapping. The objective here is to minimize the bounds on the generalization error of a model on unseen data before rather than minimizing the mean-square error over the training dataset itself (data fitting). Consequently, the optimization problem could be formulated as minimizing the following cost function:

$$L(\mathbf{w}, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i, \tag{3}$$

subject to the constraint:

$$\begin{gathered} y_i\left(\mathbf{w}^T\Phi(\mathbf{x}_i) + b\right) \geq 1 - \zeta_i, \quad i = 1, 2, \ldots, n, \\ \zeta_i \geq 0 \quad \forall i, \end{gathered} \tag{4}$$

where $\mathbf{w}$ is a weighting vector and $\Phi(\cdot)$ is a nonlinear mapping function. The ζ_i represents the tolerance error allowed for each sample to be on the wrong side of the margin (called hinge loss). Note that minimization of the first term in (3) increases the separation (margin) between the two classes, whereas minimization of the second term improves fitting accuracy. The tradeoff between complexity (or margin separation) and fitting error is controlled by the regularization parameter C.

It stands to reason that such a nonlinear formulation would suffer from the curse of dimensionality (i.e., the dimensions of the problem become too large to solve) [26, 27]. However, computational efficiency is achieved from solving the dual optimization problem instead of (3). The dual optimization problem is convex but positive-semidefinite (global but not necessarily unique solution). However, the complexity in this case is dependent only on the number of samples and not on the dimensionality of the feature space. Moreover, because of its rigorous mathematical foundations, it overcomes the "black box" stigma of other learning methods such as neural networks. The prediction function in this case is characterized by only a subset of the training data known as support vectors $\mathbf{s}_i$:

$$f(\mathbf{x}) = \sum_{i=1}^{n_s}\alpha_i y_i K(\mathbf{s}_i, \mathbf{x}) + \alpha_0, \tag{5}$$

where n_s is the number of support vectors, α_i are the dual coefficients determined by quadratic programming, and $K(\cdot, \cdot)$ is the kernel function. Typical kernels include

$$\begin{gathered} \text{Polynomials: } K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}' + c)^q \\ \text{Radial basis function (RBF): } K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}'\|^2\right), \end{gathered} \tag{6}$$

where c is a constant, q is the order of the polynomial, and σ is the width of the radial basis functions. Note that the kernel in these cases acts as a similarity function between sample points in the feature space. Moreover, kernels enjoy closure properties, that is, one can create admissible composite kernels by weighted addition and multiplication of elementary kernels. This flexibility allows for the construction of a neural network by using a combination of sigmoidal kernels. Alternatively, one could choose a logistic regression equivalent kernel by replacing the hinge loss with the binomial deviance.

2.4. Model Variable Selection. Multivariate analysis often involves a large number of variables or features [28]. The main features that characterize the observations are usually unknown. To address this, dimensionality reduction or subset selection aims to find the "significant" set of features. Although an ideal method would marginalize redundant variables, such variables usually complicate data exploration without significance. As a result, identifying the best subset of features is a challenge, especially in the case of nonlinear models. The objective remains to reduce the model complexity, decrease the computational burden, and improve the generalizability on unseen data.

In any given pattern recognition problem, there is a large number, K, of possible modeling features that could be extracted from the patients' data, making it necessary to select a finite set of features d that has the most discriminating power for the problem. An optimal subset would be determined by an exhaustive search, which would yield $\binom{K}{d}$. Fortunately, there are other and more efficient alternatives [29]. The straightforward method is to make an educated guess based on experience and domain knowledge, then apply a feature transformation (e.g., principle component analysis (PCA)) [29, 30]. It is also common to apply sensitivity analysis by using an organized search such as sequential forward selection, sequential backward selection, or a combination of both [29]. Different methods for sensitivity analysis have been proposed in literature; one such proposal is to monitor the increment in the training error when a feature is replaced by its mean. The feature is considered relevant if the increment is high. A recursive elimination technique that is based on machine learning has been also suggested [31]. In this case, the dataset is initialized to contain the whole set, the predictor (e.g., SVM classifier) is trained on the data, the features are ranked according to a certain criteria (e.g.,$\|\mathbf{w}\|$), and iteration continues by eliminating the lowest ranked feature. In our previous work [10], we used model-order determination based on information theory and resampling techniques to select the significant variables.

2.5. Evaluation and Validation Methods. To evaluate the performance of our classifiers, we used Matthew's correlation coefficient (MCC) [32] as a performance evaluation metric for classification. An MCC value of 1 would indicate perfect classification, a value of -1 would indicate anticlassification, and a value close to zero would indicate no correlation. The value of this metric, however, is proportional to the area under the receiver-operating characteristics (ROCs) curve. For ranking evaluation, we used Spearman's correlation, which provides a robust estimator of trend. This is a desirable property, particularly when ranking the quality of treatment plans for different patients.

We used resampling methods (leave-one-out cross-validation (LOO) and bootstrap) for model selection and performance comparison purposes. These methods provide statistically sound results when the available data set is limited [33]. Application of these methods for radiotherapy outcome modeling is reviewed in our previous work [10].

2.6. Visualization of Higher Dimensional Data. Prior to applying a kernel-based method, it is informative to run a screening test by visualizing the data distribution. This requires projecting the data into a lower dimensional space. Techniques such as principal component analysis (PCA) and multidimensional scaling (MDS) allow visualization of complex data in plots with reduced dimensions, often two- or three-dimensional spaces [34]. In PCA analysis, the principal components (PCs) of a data matrix $\mathbf{X}$ (with zero mean) are given by

$$\mathrm{PC} = U^T\mathbf{X} = \Sigma V^T, \tag{7}$$

where $U\Sigma V^T$ is the singular value decomposition of $\mathbf{X}$. This is equivalent to transformation into a new coordinate system such that the greatest variance by any projection of the data would lie on the first coordinate (first PC), the second greatest variance on the second coordinate (second PC), and so on.

MDS provides a nonlinear mapping that approximates local geometric relationships between points in high-dimensional space on a low-dimensional space that can be visualized. The objective function referred to here as the stress could be written as

$$L\left(y_1, y_2, \dots, y_n\right) = \sum_{i<j} \left(d_{ij} - \delta_{ij}\right)^2, \tag{8}$$

where δ_{ij} represents the target lower-dimensional distances and d_{ij} represents higher dimensional distances of the points with K features each. The optimization problem in (8) is solved as a nonlinear least squares problem using the standard Levenberg-Marquardt algorithm.

2.7. 3-Way Experimental Design for Predicting RP from Proteomic Data. The design of our prospective study utilized tools offered within Rosetta software extensively. Four different treatment groups were identified to the program: (1) control pretreatment (control-pre); (2) control posttreatment (control-post); (3) disease pretreatment (disease-pre); (4) disease posttreatment (disease-post). For these four sets of MS data (generated from four serum samples), we used the default parameters of Rosetta Elucidator (Rosetta Inpharmatics LLC, Seattle, Wash, USA) to convert raw data into aligned, combined, and ratio data as described briefly below. Annotations from peptides with Ion Scores >40 were applied to all corresponding features. Functional analysis of the identified proteins was carried using the MetaCore software (GeneGo Inc., St Joseph, Mich, USA).

OverView of Mass Spectroscopy Analysis. The Rosetta Elucidator uses raw mass spectroscopy (MS) data as an input and applies multiple normalizations and transformations in order to align, quantify, and compare features between samples. The steps of this process calculate three different types of data from the raw spectral input: aligned data, combined data, and ratio data. Aligned data have been converted into peak regions, or features, with corresponding

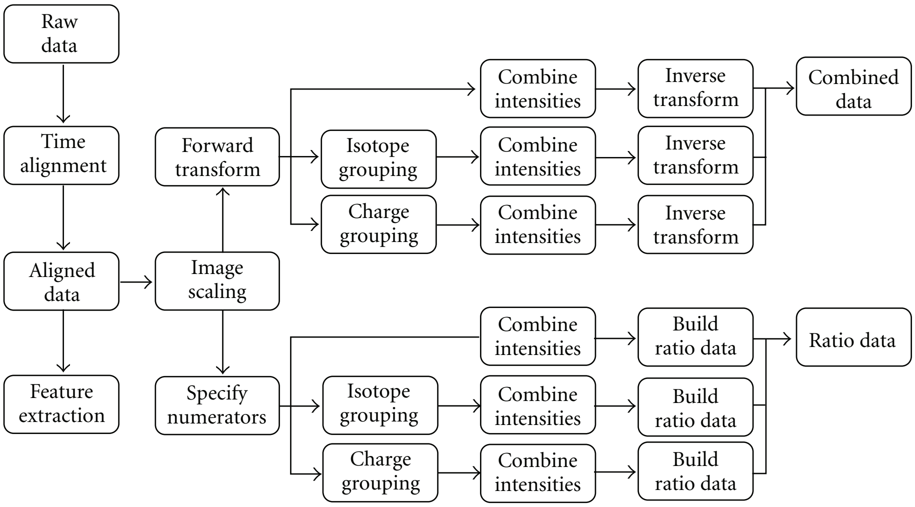


Figure 5: Diagram describing the general data processing steps Rosetta Elucidator use to calculate different data types out of raw, spectral input from a mass spectrometer (from Rosetta Inpharmatics LLC, Seattle, Wash, USA).

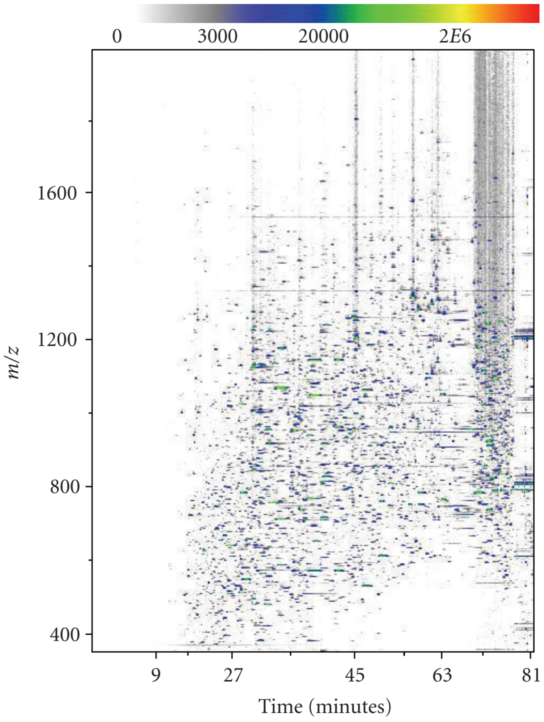


Figure 6: A 2D depiction of the raw MS data from the control-pre sample. The graph plots *m/z* versus elution time and displays intensity at a given point with color.

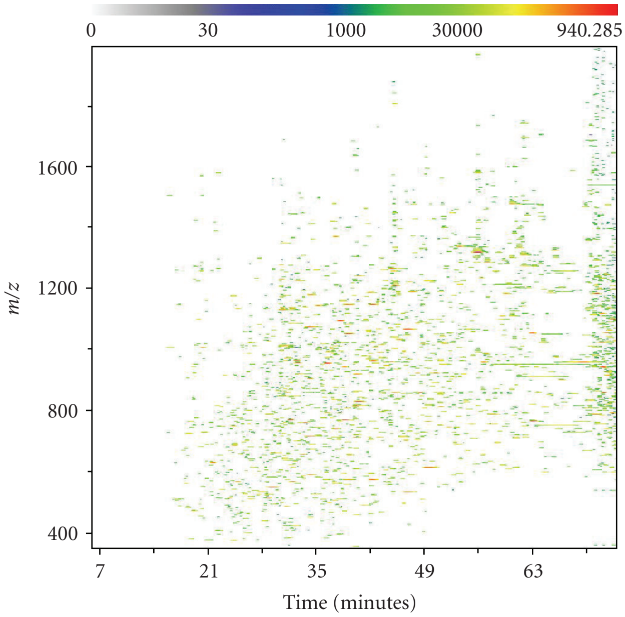


Figure 7: A representation of aligned data features generated by Rosetta Elucidator from the control-pre mass spectrum. The *m/z* measurement from the mass spectrometer is plotted against the liquid chromatography elution time, with a scale of color depicting the intensity (total ion current or TIC) measured at each point. The Elucidator system defines features by mathematically identifying local intensity peak regions against background noise.

intensity values that can be compared across samples. Combined data are composed of features with intensity values scaled by global mean intensities and transformed to stabilize error variance across samples. Ratio data are calculated through scaled intensity comparison between any two given sets of aligned data. The process is summarized in Figure 5 and described in the following.

Data Alignment. In its first stages, the Elucidator program transforms raw data into aligned data. Since peaks are not initially defined in the data, alignment starts at the level of the spectrum. The raw data for each sample include extremely precise mass to charge ratios (*m/z* ratios), times of elution from the liquid chromatogram, and detected intensity values for all ionized protein fragments. These

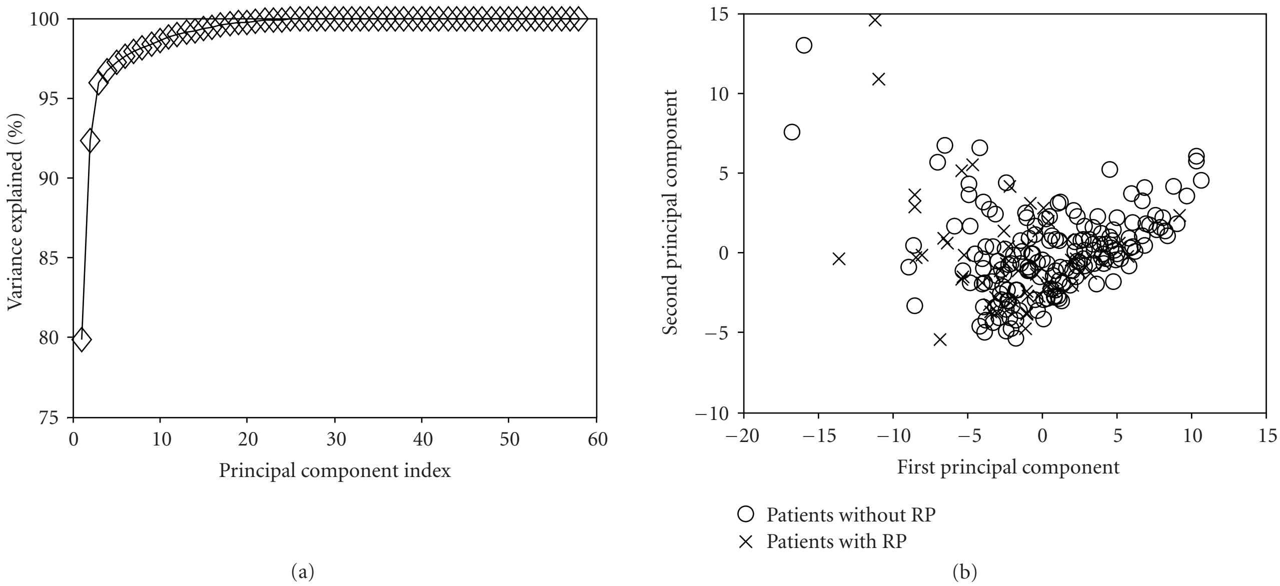


Figure 8: Visualization of the 58 variables related to RP by PCA. (a) Variation explanation versus principle component (PC) index. This represents the variance of the data model about the mean prognostic input factor values. (b) Data projection into the first two components space. Note the high overlap in projection space, which suggests nonlinear kernel modeling to achieve better predictive power.

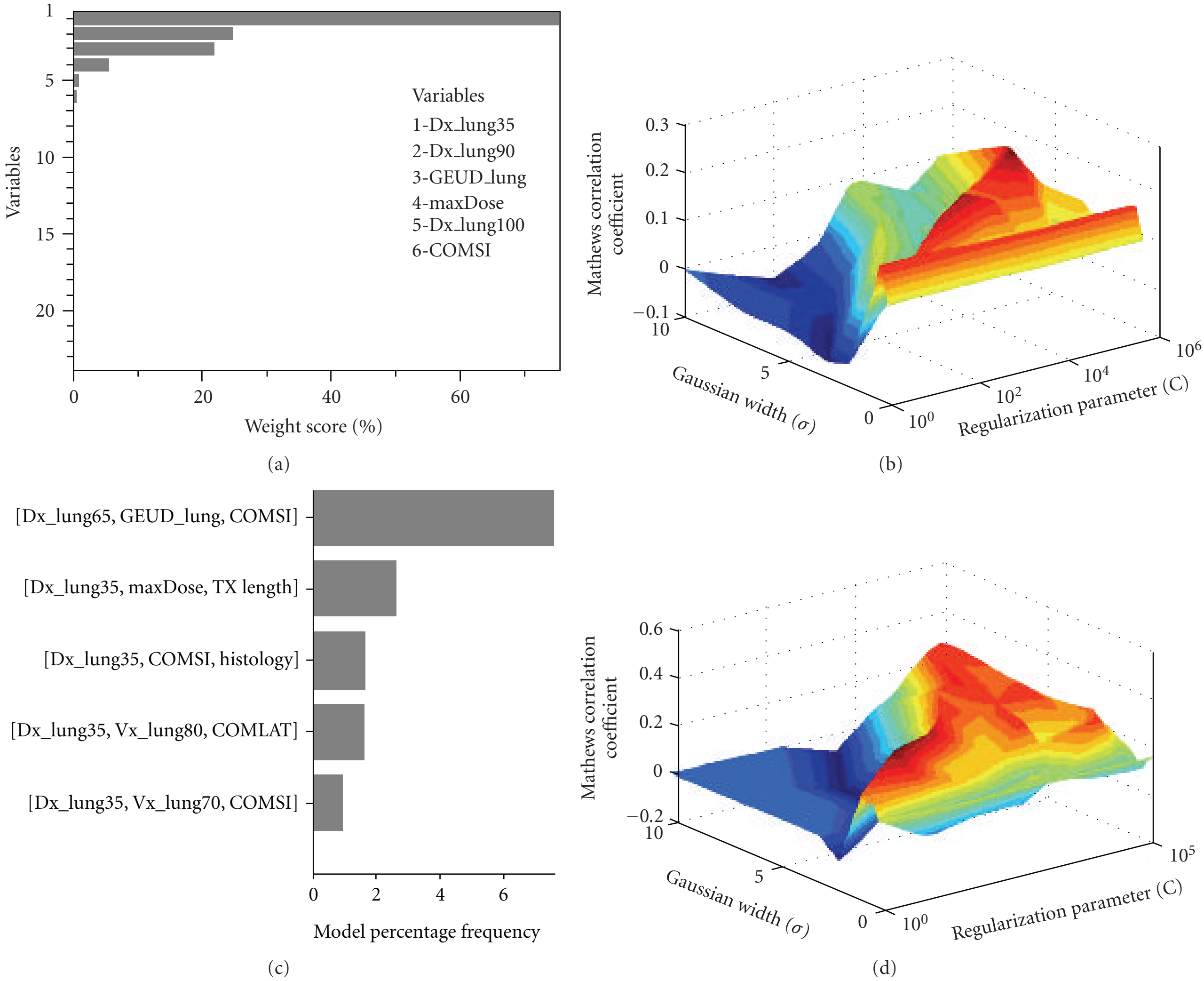
values are converted into a pixelated image with an *m/z* axis, an elution time axis, and corresponding intensity values visualized with pixel color (Figure 6). From these raw MS images, a master image is chosen and all remaining raw images are aligned to that common spectrum. The main purpose of initial spectral alignment is to correct for variations in elution time that occur between MS runs. Shifting a spectrum in time to match a master image allows for meaningful comparison between corresponding peaks in different samples. Once this time-alignment has been executed, the noise and background of each image are removed to generate aligned data that can be viewed in the system.

Feature Extraction. To extract meaningful peak regions, or features, from aligned data, a merged image is created from all the aligned images of the samples. To accomplish this, intensity values are averaged within treatment groups at each *m/z* and charge point. The resulting averaged treatment images are then averaged again across all treatments to generate a global merged image. Features can then be defined by overlaying ellipses or other two-dimensional shapes, called masks, to capture appropriate peak regions. The result across an experiment is a set of unique features with intensities measured by total ion current (TIC). Each individual feature represents a single isotopic mass peak from one of the charge states of a single peptide in a sample. Following feature extraction, the features can be grouped by isotope and the resulting isotope groups can be grouped by charge in order to capture all the features corresponding to a single peptide. An example of aligned data with extracted features is shown in Figure 7.

Combined Data. Despite this extensive process, aligned data generated by the Rosetta Elucidator system is still not the most appropriate for the comparative questions we are addressing. Aligned data generated from multiple samples does not correct for certain experimental errors and variations that occur between runs. In order to generate the most meaningful data for comparison across samples, Rosetta Elucidator converts aligned data into combined data. The first step in this transformation is a form of intensity scaling that uses the mean intensity (or brightness) of a sample, possibly the mean average brightness of samples in a treatment group, and the mean average brightness across an entire experiment. The mean brightness of a sample is calculated by excluding any missing values and then averaging the lowest 90% of feature intensity values. Each intensity value is normalized by the mean intensity of its treatment condition and the global mean intensity across the experiment. This ensures that samples and treatments share a common mean intensity, further facilitating comparisons at the level of features, isotope groups, or charge groups. Following intensity scaling, the Elucidator system applies an error model-based transformation to stabilize the noise variance over the range of intensities in use. The transform function, shown below, converts the noise variance across all samples to a constant value:

$$\hat{x} = \frac{\ln\left(b^2 + 2a^2 \cdot x + 2\sqrt{c^2 + b^2 \cdot x + a^2 \cdot x^2}\right)}{a} + d, \qquad (9)$$

where the a and b terms are related to the type of MS technology used. The a term is related to the fraction error of the instrument and the b term is related to the Poisson error of the instrument. In our experiment, we used a Linear Trap Quadrupole Orbitrab (LTQ-ORBI) mass spectrometer, which has a fraction error of 0.05 and a



(a) (b)

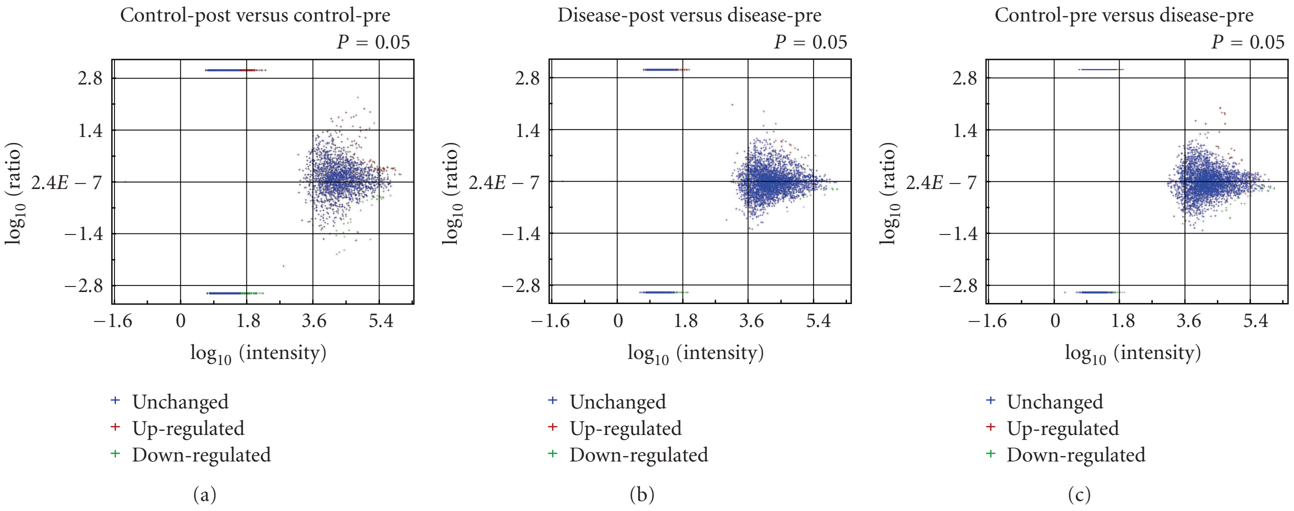
(c) (d)

Figure 9: RP with a premodeling variable selection using (a), (b) the recursive feature elimination (RFE) method. Variables were chosen from a pool of 58 dosimetric, positional, and clinical variables. The top 23 variables selected by SVM-RFE are shown after applying a pruning step to correct for multicollinearity (RS = 0.75)(RS = 0.75). The top 6 variables (by applying a cutoff of 5% weighting score) were used for modeling pneumonitis. (b) An SVM-RBF classifier was tested on LOO data. (c), (d) Multimetric logistic regression approach. (c) The frequency of selected models order of 3 using our two-step resampling methods. The best-selected model consisted of three parameters (D35, COM-SI, and maximum dose). (d) The results of applying the SVM methodology with RBF kernels using these selected variables on LOO testing data. Note the improved performance in this case compared to RFE variable selection.

Poisson error of 15 000. The c term depends upon each feature's background value, which is an error model output for aligned data that calculates the background intensity surrounding the feature (ideally zero). An average of the background value is calculated over all features i and all treatments j in the experiment. The term d is related through a logarithm transform to a, b, and c. Following this forward transform, the transformed intensity values are averaged across all samples in the experiment to generate a separate combined intensity value. This combined intensity value is set apart from the individual sample intensity values and is calculated for later comparative and testing purposes. To generate the final combined data set, all intensities (including the combined intensity) must undergo an inverse transformation.

Ratio Data. A final type of data, called ratio data, is calculated from two input sets of aligned data, one marked as a numerator and the other marked as a denominator. Ratio data is especially informative for our experiment because it provides a way to analyze relative intensity changes that occur across the same feature in different treatment groups as discussed below.

Feature Annotation. With aligned data, combined data, and ratio data calculated automatically as part of our experimental design within Rosetta, we proceeded to annotate the sample features with the initial MS/MS peptide and protein identifications. All peptides with an Ion Score greater than 40, as calculated in Mascot search engine for peptide



(a) (b) (c)

Figure 10: Categorization of upregulated and downregulated features for the ratio data as a function of average intensity for (a) control-post to control-pre, (b) Disease-post to disease-pre, and (c) control-pre to disease-pre.

identification (Matrix Science Ltd., Boston, Mass, USA) were associated with their corresponding feature in Rosetta Elucidator.

3. Experimental Results

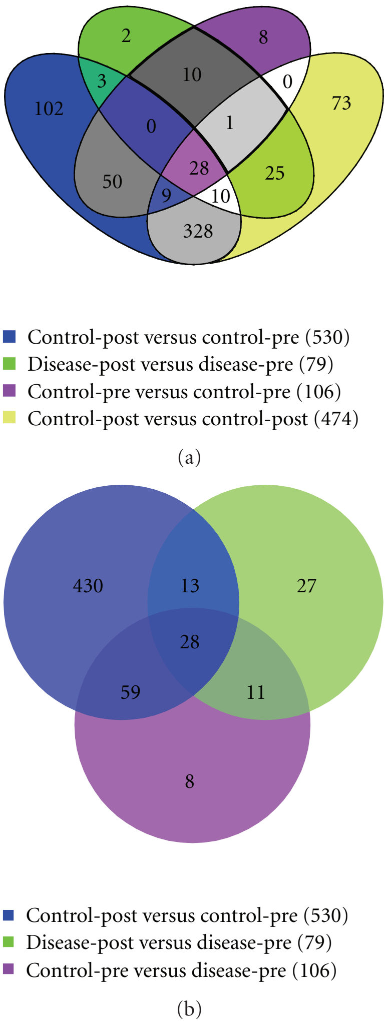
3.1. Dose-Volume RP Model

Data Exploration. In Figure 8, we present PCA analysis of RP, with a pool of 58 variables. This pool included clinical variables (age, gender, race, chemo, stage, histology, treatment, etc.), dosimetric variables, such as V_x (volume getting at least x Gy), D_x (minimum dose to the hottest x% volume), and the relative location of the tumor within the lung. Notice that more than 93% of the variations in the input data were explained by the first two components (Figure 8(a)). Additionally, the overlap between patients with and without radiation pneumonitis is very high (Figure 8(b)), suggesting that there is no linear classifier that can adequately separate these two classes.

Kernel-Based Modeling. We first explored the effect of variable selection over the entire variable pool on the prediction of pneumonitis in the lung using support vector machine with a radial basis function kernel (SVM-RBF) as a classifier. In Figure 9(a), we show the top 30 selected variables using a recursive-feature-elimination SVM method, which was previously shown to be an excellent method for gene selection in microarray studies [31]. We used variable pruning to account for multicolinearity of correlated variables in this case. In Figure 9(b), we show the resulting SVM-RBF classifier using the top six variables (using a cutoff of 5% weighting score). The best MCC obtained was 0.22. In Figure 9(c), we show the results of variable selection using our previous multimetric approach based on model order selection and resampling with logistic regression [10, 19]. The model order was determined to be 3 with variables of D35, max dose, and COM-SI (center-of-mass of tumor location in the superior inferior direction) [35]. Figure 9(d) shows the evaluation results of applying the SVM methodology with RBF kernels using these selected variables. The resulting correlation (MCC = 0.34) on LOO testing data significantly improved our previously achieved multimetric logistic regression by 46%. The basic interpretation of this improvement is that the SVM automatically identified and accounted for interactions between the model variables. Despite the improvement, the model still does not achieve correlations levels that could be applied with high confidence in clinical practice. This is possibly because the model is unable to account for biological effects adequately, which we might need to incorporate as analyzed next.

3.2. Proteomic Identification of RP. Using the 3-way methodology described in Section 2.7, we identified a group of features associated with RP by overlaying multiple subgroups of ratio data as follows. First, we organized subgroups of ratio data that displayed significant intensity changes between any two samples of interest. Significance was determined based on the P-value of each feature in a given set of ratio data. A P-value less than .05 was used as a cutoff. In this step, 11 979 unique features were identified after spectral alignment across the four samples. Of these 458 features directly matched, a peptide with an Ion Score >40 and 1289 features were annotated when direct peptide matches (with Ion Scores >40) were applied to all features in the same isotope group. Significant features could be further divided into upregulated and downregulated categories based on the sign of the fold change as shown in Figure 10.

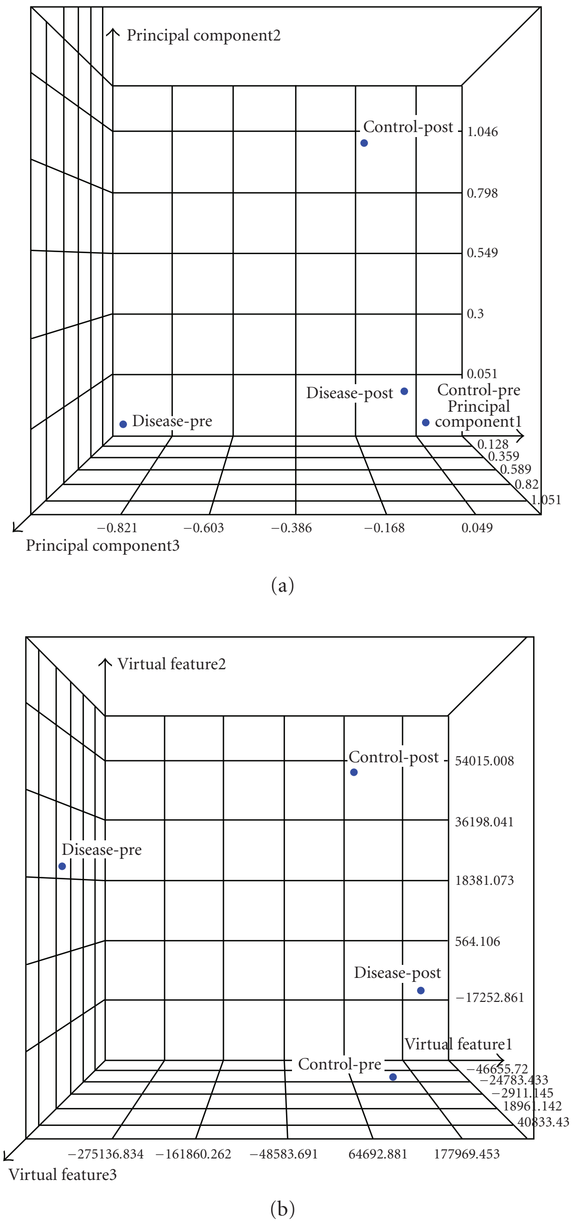
Secondly, features that significantly changed intensity between control-pre and control-post were overlaid with significant features that changed between disease-pre and disease-post. Shared features between these two datasets indicated candidate peptides that changed expression due to radiation. Alternatively, features unique to the disease-post



(a)

(b)

Figure 11: Diagram depicting the shared features between different sets of ratio data. The features in an individual set are those that displayed a significant change in intensity between the two members of the ratio. (a) All four samples, (b) the three ratios used to extract the RP candidates from the overlaid: control-pre to control-post, disease-pre to disease-post, and control-pre to disease-pre. Eleven features uniquely associated to a hypersensitive reaction as well as differential between patients before treatment.



(a)

(b)

Figure 12: Visualization of the candidate 11 features for RP (a) PCA and (b) MDS. Note the separation between control-pre and disease-pre and disease-post and disease-pre as anticipated from the experimental design strategy we followed to extract these features.

versus disease-pre significant dataset were considered associated with a deleterious, hypersensitive reaction to radiation therapy, RP in our case. Using this hypersensitive dataset, we then overlaid the significant features from control-pre versus disease-pre. The features shared between these datasets are not only associated with RP, but also can be detected (due to differential concentrations) before treatment initiates. The results of these comparisons are summarized in the Venn diagrams of Figure 11.

As noted from Figure 11(b), 41 features significantly changed after treatment in both patients. This can be attributed to regular radiation response. In addition, there were 489 significant features that were uniquely associated with the control case and 38 that were uniquely associated with the disease case. Eleven features were uniquely associated with a hypersensitive reaction as well as differential expression between patients before treatment, which represent our RP candidates. The relationship between these features and the original samples is represented in the PCA and MDS analyses of Figure 12. It is noticed that the separation between control-pre and disease-pre and disease-post and disease-pre is as anticipated from the experimental design strategy we followed to extract these features.

These 11 features were annotated as described in Section 2.7 and four proteins were identified as potential biomarkers for RP. All the identified proteins were downregulated postradiotherapy treatment and were known to play roles in inflammation responses. Two of these protein families were related to tissue remodeling, cognitive disorders, and fibrosis; one protein was part of the angiotensin-renin system, and the last protein seems to play a role in cytokine expression (interleukins and tumor necrosis factor).

4. Discussion

Modeling of radiotherapy outcomes constitutes a challenging problem due to the complex interaction between physical and biological factors. Better understanding of these relationships and the ability to develop predictive models of patients' treatment outcomes would lead to personalized treatment regimens. The tremendous increase in patient-specific clinical and biological information in conjunction with developing proper datamining methods and bioinformatics tools could potentially revolutionize the century old concepts of radiobiology and potentially improve the quality of care for radiation oncology patients.

In this work, we presented our methodology for making use of currently existing treatment planning archives to develop dose-volume models. We have demonstrated that supervised machine learning methods based on nonlinear kernels could be used to improve prediction of RP by a factor of 46% compared to traditional logistic regression methods. Potential benefits of these methods could be assessed based on PCA analysis of this data, where nonlinear kernels could be applied to resolve overlapping classes by mapping to higher-dimensional space [36]. We have applied resampling methods based on LOO to assess generalizabilty to unseen data and avoid overfitting pitfalls. Despite the gain in performance we attained from kernel methods, our results show that the best predictive model of RP has an MCC of 0.34 on LOO suggesting that our current variable space of clinical and physical dosimetric variables may not be adequate to describe the observed outcomes. This is despite the inclusion of high-order interaction terms using the SVM machinery. Therefore, we are currently exploring the inclusion of biological variables from peripheral blood draws to improve the prediction power of our RP model. Toward this goal, we have proposed a prospective study that builds upon our earlier retrospective analysis to delineate dose-volume effects in the onset of RP and include "missing" biological variables from minimally invasive clinical procedures inoperable NSCLC patients.

We have conducted a proteomic analysis of blood serum samples. Specifically, we have proposed a 3-way design strategy in order to distinguish between patient's variations, confounding radiation effects, and hypersensitivity predictors using intensity ratio changes. To test the validity of our design, PCA and MDS plots were used to measure separation between the samples in the estimated feature space. Our proteomic analysis was based on data from only two samples, but the results still provided promising candidates to validate with biochemical assays in a larger cohort. The entire study size of nineteen patients is an arguably small sample size as well, but according to our current protocol the number of patients in this study will increase every year, as new patients are recruited, with a final goal of 100–120 patients participating. Ongoing generation and validation of candidate proteins through additional mass spectrometry runs and extensive biochemical assays should provide increasingly interesting and accurate candidate proteins. Our feature selection strategy for candidate proteins is simplistic at this point, but we plan to make effective use of new emerging methodologies in statistical analysis of such data [37–40]. However, further investigation of datamining approaches to extract proper features and identify corresponding proteins with higher confidence from limited datasets is still required.

In our future work, we plan to further validate the derived proteins by examining their functional role by querying protein databases and measure their expression using Enzyme-Linked ImmunoSorbent Assay (ELISA). If successful, this data would be mixed with the developed dose-volume model using SVM-RBF and we will test the overall prediction on prospective data. Thus, we would be able to benefit from both retrospective and prospective data in our model building strategy.

5. Conclusions

We have demonstrated machine-learning application and a proteomics design strategy for building a predictive model of RP. The machine learning methods efficiently and effectively handle high-dimensional space of potentially critical features. We have applied this model successfully to interrogate dose-volume metrics. Our proteomics strategy seems to identify relevant biomarkers to inflammation response. Furthermore, we are currently investigating incorporation of these biomarkers into our existing dose-volume model of RP to improve its prediction power and potentially demonstrate its feasibility for individualization of radiotherapy of NSCLC patients.

Acknowledgments

The authors would like to thank Dr. Reid Townsend for his assistance in the proteomic analysis. This work was partially supported by NIH Grant 5K25CA128809-02.

References

[1] American Cancer Society, *Cancer Facts and Figures*, American Cancer Society, Atlanta, Ga, USA, 2008.

[2] J. M. Balter and Y. Cao, "Advanced technologies in image-guided radiation therapy," *Seminars in Radiation Oncology*, vol. 17, no. 4, pp. 293–297, 2007.

[3] S. Webb, *The Physics of Three Dimensional Radiation Therapy: Conformal Radiotherapy, Radiosurgery and Treatment Planning*, Institute of Physics, Bristol, UK, 2001.

[4] C. M. L. West, R. M. Elliott, and N. G. Burnet, "The genomics revolution and radiotherapy," *Clinical Oncology*, vol. 19, no. 6, pp. 470–480, 2007.

[5] B. G. Wouters, "Proteomics: methodologies and applications in oncology," *Seminars in Radiation Oncology*, vol. 18, no. 2, pp. 115–125, 2008.

[6] D. E. Lea, *Actions of Radiations on Living Cells*, Cambridge University Press, Cambridge, UK, 1946.

[7] E. J. Hall and A. J. Giaccia, *Radiobiology for the Radiologist*, Lippincott Williams & Wilkins, Philadelphia, Pa, USA, 6th edition, 2006.

[8] J. P. Kirkpatrick, J. J. Meyer, and L. B. Marks, "The linear-quadratic model is inappropriate to model high dose per fraction effects in radiosurgery," *Seminars in Radiation Oncology*, vol. 18, no. 4, pp. 240–243, 2008.

[9] S. Levegrün, A. Jackson, M. J. Zelefsky, et al., "Analysis of biopsy outcome after three-dimensional conformal radiation therapy of prostate cancer using dose-distribution variables and tumor control probability models," *International Journal of Radiation Oncology, Biology, Physics*, vol. 47, no. 5, pp. 1245–1260, 2000.

[10] I. El Naqa, J. D. Bradley, A. I. Blanco, et al., "Multivariable modeling of radiotherapy outcomes, including dose-volume and clinical factors," *International Journal of Radiation Oncology, Biology, Physics*, vol. 64, no. 4, pp. 1275–1286, 2006.

[11] S. Levegrün, A. Jackson, M. J. Zelefsky, et al., "Fitting tumor control probability models to biopsy outcome after three-dimensional conformal radiation therapy of prostate cancer: pitfalls in deducing radiobiologic parameters for tumors from clinical data," *International Journal of Radiation Oncology, Biology, Physics*, vol. 51, no. 4, pp. 1064–1080, 2001.

[12] L. B. Marks, "Dosimetric predictors of radiation-induced lung injury," *International Journal of Radiation Oncology, Biology, Physics*, vol. 54, no. 2, pp. 313–316, 2002.

[13] J. O. Deasy, A. Niemierko, D. Herbert, et al., "Methodological issues in radiation dose-volume outcome analyses: summary of a joint AAPM/NIH workshop," *Medical Physics*, vol. 29, no. 9, pp. 2109–2127, 2002.

[14] S. M. Bentzen, "From cellular to high-throughput predictive assays in radiation oncology: challenges and opportunities," *Seminars in Radiation Oncology*, vol. 18, no. 2, pp. 75–88, 2008.

[15] F.-M. Kong, R. Ten Haken, A. Eisbruch, and T. S. Lawrence, "Non-small cell lung cancer therapy-related pulmonary toxicity: an update on radiation pneumonitis and fibrosis," *Seminars in Oncology*, vol. 32, supplement 3, pp. S42–S54, 2005.

[16] Y. Chen, P. Okunieff, and S. A. Ahrendt, "Translational research in lung cancer," *Seminars in Surgical Oncology*, vol. 21, no. 3, pp. 205–219, 2003.

[17] J. O. Deasy, A. I. Blanco, and V. H. Clark, "CERR: a computational environment for radiotherapy research," *Medical Physics*, vol. 30, no. 5, pp. 979–985, 2003.

[18] I. El Naqa, G. Suneja, P. E. Lindsay, et al., "Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose-volume outcome relationships," *Physics in Medicine and Biology*, vol. 51, no. 22, pp. 5719–5735, 2006.

[19] J. O. Deasy and I. El Naqa, "Image-based modeling of normal tissue complication probability for radiation therapy," in *Radiation Oncology Advances*, M. Mehta and S. Bentzen, Eds., pp. 211–252, Springer, New York, NY, USA, 2008.

[20] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[21] I. El Naqa, J. D. Bradley, and J. O. Deasy, "Machine learning methods for radiobiological outcome modeling," in *Proceedings of the AAPM Symposium on Physical, Chemical and Biological Targeting in Radiation Oncology*, M. Mehta, B. Paliwal, and S. Bentzen, Eds., vol. 14, pp. 150–159, Medical Physics, Seattle, Wash, USA, July 2005.

[22] I. El-Naqa, Y. Yang, N. P. Galatsanos, R. M. Nishikawa, and M. N. Wernick, "A similarity learning approach to content-based image retrieval: application to digital mammography," *IEEE Transactions on Medical Imaging*, vol. 23, no. 10, pp. 1233–1244, 2004.

[23] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Transactions on Medical Imaging*, vol. 21, no. 12, pp. 1552–1563, 2002.

[24] B. Schèolkopf, K. Tsuda, and J.-P. Vert, *Kernel Methods in Computational Biology*, MIT Press, Cambridge, Mass, USA, 2004.

[25] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, UK, 2004.

[26] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-Color Illustrations*, Springer, New York, NY, USA, 2001.

[27] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice-Hall, Englewood Cliffs, NJ, USA, 2nd edition, 1999.

[28] I. Guyon and A. Elissee, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[29] R. Kennedy, Y. Lee, B. van Roy, C. D. Reed, and R. P. Lippman, *Solving Data Mining Problems through Pattern Recognition*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1998.

[30] L. A. Dawson, M. Biersack, G. Lockwood, A. Eisbruch, T. S. Lawrence, and R. K. Ten Haken, "Use of principal component analysis to evaluate the partial organ tolerance of normal tissues to radiation," *International Journal of Radiation Oncology, Biology, Physics*, vol. 62, no. 3, pp. 829–837, 2005.

[31] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[32] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975.

[33] P. I. Good, *Resampling Methods: A Practical Guide to Data Analysis*, Birkhäuser, Boston, Mass, USA, 3rd edition, 2006.

[34] W. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, Springer, Berlin, Germany, 2003.

[35] A. J. Hope, P. E. Lindsay, I. El Naqa, et al., "Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters," *International Journal of Radiation Oncology, Biology, Physics*, vol. 65, no. 1, pp. 112–124, 2006.

[36] I. El Naqa, J. D. Bradley, and J. O. Deasy, "Nonlinear kernel-based approaches for predicting normal tissue toxicities," in *Proceedings of the 7th International Conference on Machine Learning and Applications (ICMLA '08)*, pp. 539–544, San Diego, Calif, USA, December 2008.

[37] H.-D. Zucht, J. Lamerz, V. Khamenia, et al., "Datamining methodology for LC-MALDI-MS based peptide profiling," *Combinatorial Chemistry & High Throughput Screening*, vol. 8, no. 8, pp. 717–723, 2005.

[38] S. Gay, P.-A. Binz, D. F. Hochstrasser, and R. D. Appel, "Peptide mass fingerprinting peak intensity prediction: extracting knowledge from spectra," *Proteomics*, vol. 2, no. 10, pp. 1374–1391, 2002.

[39] A. I. Nesvizhskii, O. Vitek, and R. Aebersold, "Analysis and validation of proteomic data generated by tandem mass

spectrometry," *Nature Methods*, vol. 4, no. 10, pp. 787–797, 2007.

[40] B. M. Broom and K.-A. Do, "Statistical methods for biomarker discovery using mass spectrometry," in *Statistical Advances in Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics*, A. Biswas, S. Datta, J. P. Fine, and M. R. Segal, Eds., pp. 465–486, Wiley-Interscience, Hoboken, NJ, USA, 2008.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 726728, 6 pages
doi:10.1155/2009/726728

Research Article

Integrative Decomposition Procedure and Kappa Statistics for the Distinguished Single Molecular Network Construction and Analysis

Lin Wang,[1] Ying Sun,[1] Minghu Jiang,[2] and Xiguang Zheng[1]

[1] Biomedical Center, School of Electronics Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] Lab of Computational Linguistics, Tsinghua University, Beijing 100084, China

Correspondence should be addressed to Lin Wang, wanglin98@tsinghua.org.cn

Received 27 October 2008; Accepted 19 February 2009

Recommended by Zhenqiu Liu

Our method concentrates on and constructs the distinguished single gene network. An integrated method was proposed based on linear programming and a decomposition procedure with integrated analysis of the significant function cluster using Kappa statistics and fuzzy heuristic clustering. We tested this method to identify ATF2 regulatory network module using data of 45 samples from the same GEO dataset. The results demonstrate the effectiveness of such integrated way in terms of developing novel prognostic markers and therapeutic targets.

1. Introduction

In the postgenomic era, with microarray technologies producing great deal of gene expression data, mining these data to get insight into biological processes at system-wide level has become a challenge for bioinformatics. On one hand, due to the complex and distribute nature of biological research, there is a great deal of methods for inferring gene regulatory networks. But all these methods focused on constructing the complicated entire network calculated from the given microarray data. The tremendous amounts of genes in those networks distribute analysts' attention, so it is hard to get any clear perception of valuable knowledge from such complicated networks, let alone further study of each single gene. On the other hand, the wide spread of knowledge over independent databases aggravates the hardness of integrating comprehensive annotation information for genes and lowers the study effectiveness. Thus, a novel method integrating both single molecular network construction and highly centralized gene-functional-annotation analysis is in demand for gene network and functional analysis.

This paper proposed an integrated method based on linear programming and a decomposition procedure with integrated analysis of the significant function cluster using Kappa statistics and fuzzy heuristic clustering. Our method concentrates on and constructs the distinguished single gene network integrated with function prediction analysis by DAVID. For the distinguished single molecular network, we did (1) control and experiment comparison, (2) identification of activation and inhibition networks, (3) construction of upstream and downstream feedback networks, and (4) functional module construction. We tested this method to identify ATF2 regulation network module using data of 45 samples from one and the same GEO dataset. The results demonstrate the effectiveness of such integrated way in terms of developing novel prognostic markers and therapeutic targets.

2. Methods

2.1. Distinguished Single Molecular Network Construction. The entire network was constructed using GRNInfer [1] and GVedit tools. GRNInfer is a novel mathematic method called gene network reconstruction (GNR) tool based on linear programming and a decomposition procedure that is used for inferring gene networks. The method theoretically ensures the derivation of the most consistent network structure with respect to all of the datasets, thereby not only significantly alleviating the problem of data scarcity but also remarkably improving the reconstruction reliability. The general solution for a single dataset is the following (1), which represents all of the possible networks:

$$J = (X' - A)U\Lambda^{-1}V^T + YV^T = \hat{J} + YV^T, \qquad (1)$$

where $J = (J_{ij})_{n\times n} = \partial f(x)/\partial x$ is an $n \times n$ Jacobian matrix or connectivity matrix, $X = (x(t_1),\dots,x(t_m))$, $A = (a(t_1),\dots,a(t_m))$, and $X' = (x'(t_1),\dots,x'(t_m))$ are all $n \times m$ matrices with $x_i'(t_j) = [x_i(t_{j+1}) - x_i(t_j)]/[t_{j+1} - t_j]$ for $i = 1,\dots,n$; $j = 1,\dots,m$. $X(t) = (x_1(t),\dots,x_n(t))^T \in R^n$, $a = (a_1,\dots,a_n)^T \in R^n$, $x_i(t)$ is the expression level (mRNA concentrations) of gene i at time instance t. $y = (y_{ij})$ is an $n \times n$ matrix, where y_{ij} is zero if $e_j \neq 0$ and is otherwise an arbitrary scalar coefficient. $\wedge^{-1} = \mathrm{diag}\,(1/e_i)$ and $1/e$ is set to be zero if $e_i = 0$. U is a unitary $m \times n$ matrix of left eigenvectors, $\wedge = \mathrm{diag}\,(e_1,\dots,e_n)$ is a diagonal $n \times n$ matrix containing the n eigenvalues, and V^T is the transpose of a unitary $n \times n$ matrix of right eigenvectors.

But the entire network is too complex to get any clear perception of such complicated relationships among those genes, let alone further study of each single gene. We constructed the distinguished single molecular network by selecting the centered gene and its directly related genes based on the entire network for further study. We take into account the effectiveness of biology study in order to concentrate on single molecular network rather than the intricate entire network. It is helpful to get intensive and deep insight of the whole network. For the distinguished single molecular network, we did (1) control and experiment comparison, (2) identification of activation and inhibition networks, (3) construction of upstream and downstream feedback networks, and (4) functional module construction.

2.2. Functional Annotation Clustering. For the function of genes that is neither determined by their sequence nor by the protein families they belong to [2], the function of those genes included in the same single molecular network should not be interpreted separately, but should be analyzed together according to the whole single molecular network. This method takes into account the network nature of biological annotation contents in order to concentrate on the larger biological picture rather than an individual gene. We used DAVID to do functional annotation clustering. It changes functional annotation analysis from term- or gene-centric to biological module-centric [2] in accordance with our network analysis aim.

The DAVID gene functional clustering tool provides typical batch annotation and gene-GO term enrichment analysis for highly throughput genes by classifying them into gene groups based on their annotation term co-occurrence [3]. DAVID uses a novel algorithm to measure relationships among the annotation terms based on the degrees of their coassociation genes to group similar annotation contents from the same or different resources into annotation groups. The grouping algorithm is based on the hypothesis that similar annotations should have similar gene members. The functional annotation clustering integrates the same techniques of Kappa statistics to measure the degree of the common genes between two annotations, and fuzzy heuristic clustering to classify the groups of similar annotations according kappa values [4, 5]. The tool also allows observation of the internal relationships of the clustered terms by comparing it to the typical linear, redundant term report, over which similar annotation terms may be distributed among many other terms.

3. Results and Discussion

We tested this method using microarrays containing 22215 genes in 40 MPM tumors and 5 normal pleural tissues from one and the same GEO datasets. We identified potential tumor molecular markers and chose the top 51 significant positive genes with normalization of log2, the minimum fold change = 3.5, delta = 1.59, and a false-discovery rate of 0% using SAM [6]. We selected activating transcription factor (ATF)-2 because it is one of the most distinguished genes in MPM. It is a member of the ATF/cyclic AMP-responsive element binding protein family of transcription factors.

3.1. Normal Tissues and Tumor Comparisons of Distinguished Single Molecular Network. We, respectively, constructed the interaction network of the above 51 genes in healthy tissues and that in tumor using GRNInfer [1] and GVedit tools and selected the ATF2-centered downstream subnetworks. With comparison of these ATF2-centered subnetworks, we can get a more clear perception of the notable differences between normal tissues and tumor, as shown in Figure 1. It appeared that ATF2 inhibits C11orf9, C18orf10, C20orf31, CALD1, CAMK2G, DDX3X, FALZ, GLS, GOLGA2, ID2, NME2, NMU, NONO, PAWR, PLOD2, PSMF1, RBMS1, RIC8A, RNF10, TEAD4, TIA1, TNPO1, unknown2, unknown3, WBSCR20C, and ZF in normal tissues, as shown in Figure 1(a). It appeared that ATF2 inhibits C11orf9, C15orf5, C18orf10, C20orf31, CAMK2G, CDR2, DDX3X, FALZ, FLJ10707, GLS, GOLGA2, ID2, KRT18, LRRC1, NME2, NMU, NONO, NSUN5, OBSL1_2, PLOD2, PLXNA1, PTOV1, RBMS1, RIC8A, RNASEH1, RNF10, TEAD4, TIA1, UCK2, USP11, and ZF, while it activates CALD1 and TFAP2C in tumor, as shown in Figure 1(b).

With comparison between the two results, notable differences can be shown clearly in order to get further perception of pathological changes in MPM. For example, ATF2 target genes appeared in ATF2 activation to CALD1, TFAP2C in MPM, as only shown in Figure 2(b). Caldesmon (CALD1) is a potential actomyosin regulatory protein found in smooth muscle and nonmuscle cells [7]. Transcription factor AP2-gamma (TFAP2C) is alternatively titled AP2. Families of related transcription factors are often expressed in the same cell lineages but at different times or sites in the developing embryo. The AP2 family appears to regulate the expression of genes required for development of tissues of ectodermal origin such as neural crest and skin [8]. AP2 may also be

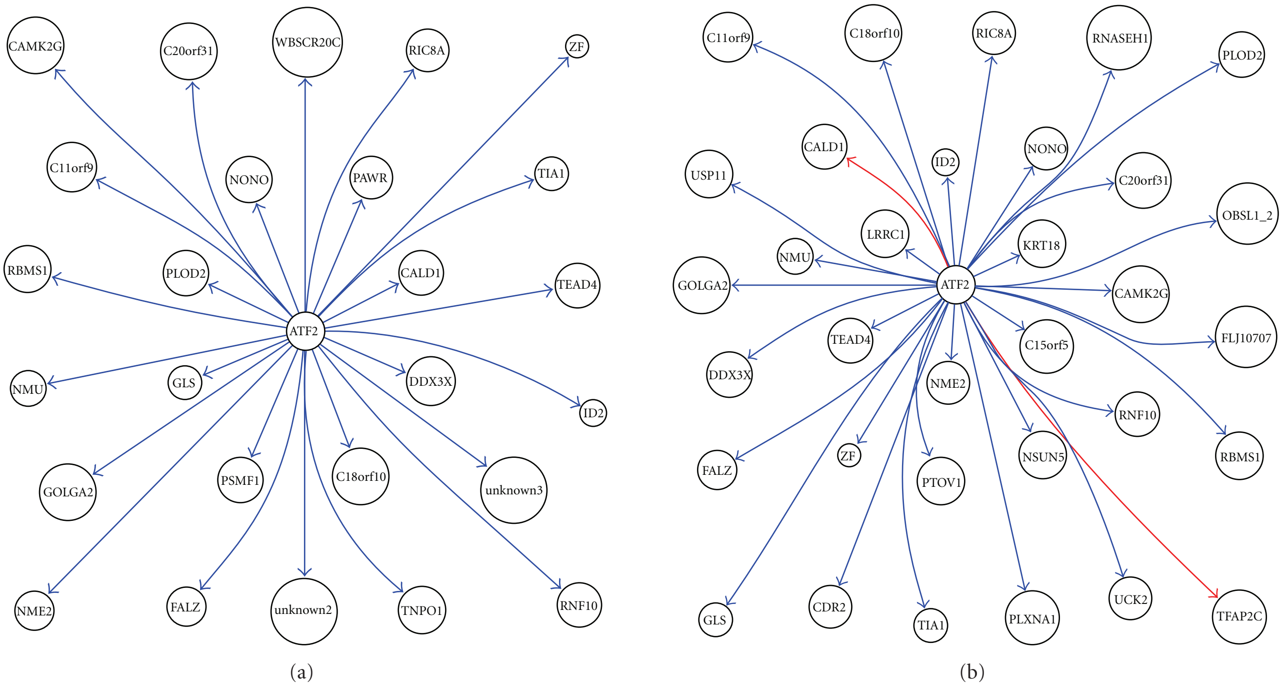


Figure 1: ATF2 downstream network in (a) normal tissue and (b) MPM tissue.

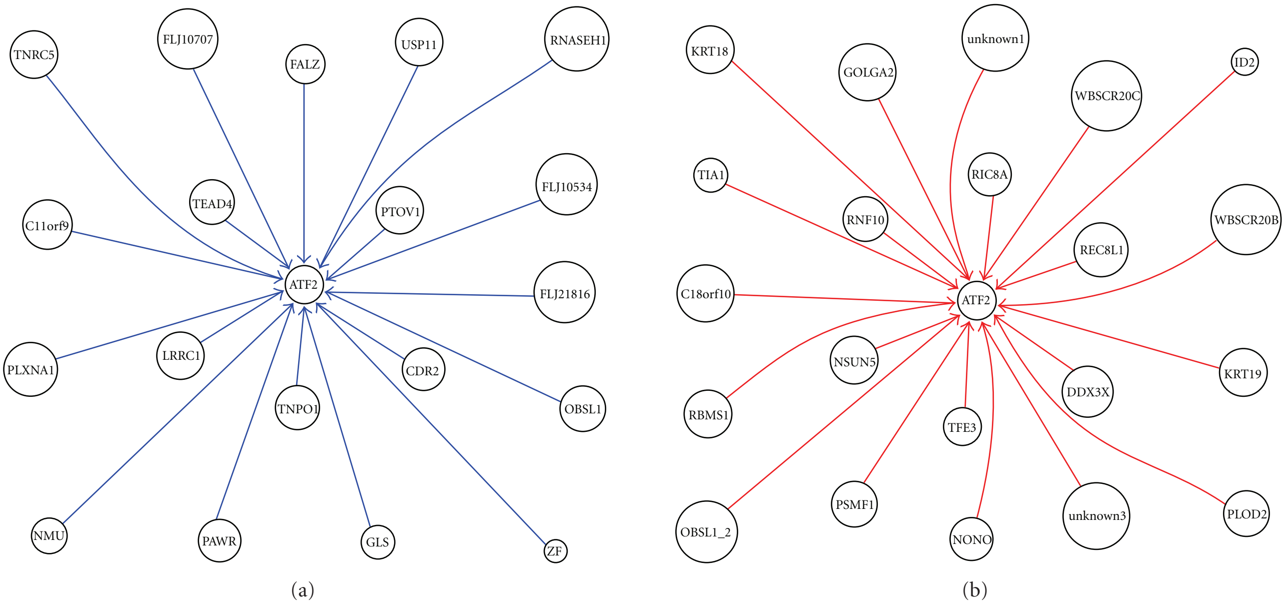


Figure 2: (a) ATF2 upstream inhibition network of MPM; (b) ATF2 upstream activation network of MPM.

involved in the overexpression of c-erbB-2 in human breast cancer cells [9].

3.2. Identification of Activation and Inhibition Networks for the Distinguished Single Molecule. We also identified the activation and inhibition networks, respectively, in order to simplify and intensify the analysis process. For example, in ATF2 upstream network of MPM, as shown in Figure 2, it appeared that C11orf9, CDR2, FALZ, FLJ10534, FLJ10707, FLJ21816, GLS, LRRC1, NMU, OBSL1, PAWR, PLXNA1, PTOV1, RNASEH1, TEAD4, TNPO1, TNRC5, USP11, and ZF inhibit ATF2, as shown in Figure 2(a), whereas C18orf10, DDX3X, GOLGA2, ID2, KRT18, KRT19, NONO, NSUN5, OBSL1_2, PLOD2, PSMF1, RBMS1, REC8L1, RIC8A, RNF10, TFE3, TIA1, unknown1, unknown3, WBSCR20B, and WBSCR20C activate ATF2, as shown in Figure 2(b).

ATF2 upstream genes TFE3, REC8L1 showed activation to ATF2. TFE3 is a member of the helix-loop-helix family

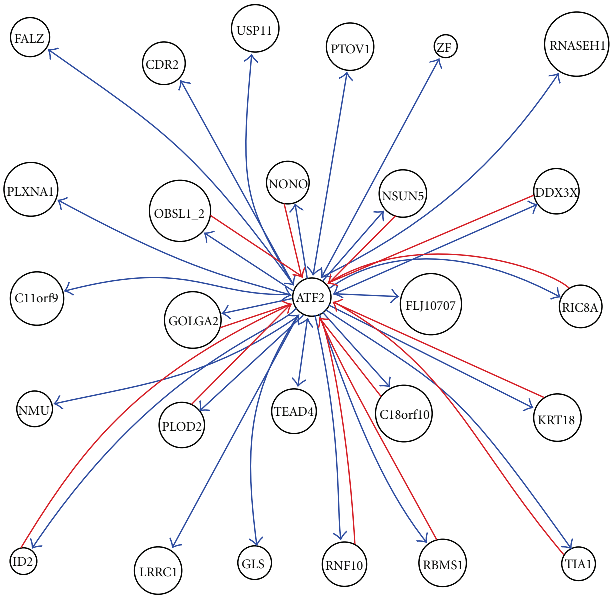


Figure 3: ATF2 feedback subnetwork of MPM.

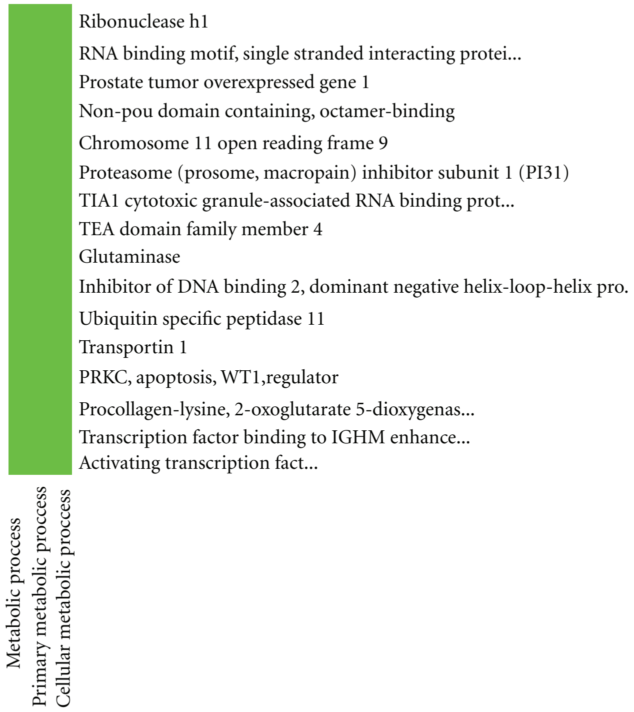


Figure 4: One ATF2 upstream gene metabolic network including RBMS1, RNASEH1, PTOV1, NONO, C11orf9, PSMF1, TIA1, TEAD4, GLS, ID2, USP11, TNPO1, PAWR, PLOD2, and TFE3.

of transcription factors and binds to the mu-E3 motif of the immunoglobulin heavy-chain enhancer and is expressed in many cell types [10]. Nakagawa et al. [11] identified TFE3 as a transactivator of metabolic genes that are regulated through an E box in their promoters which led to metabolic consequences such as activation of glycogen and protein synthesis, but not lipogenesis, in liver [11]. REC8L1 is the human homolog of yeast Rec8, a meiosis-specific phosphoprotein involved in recombination events [12]. Brar et al. (2006) showed that phosphorylation of the cohesin subunit REC8 contributes to stepwise cohesin removal [13].

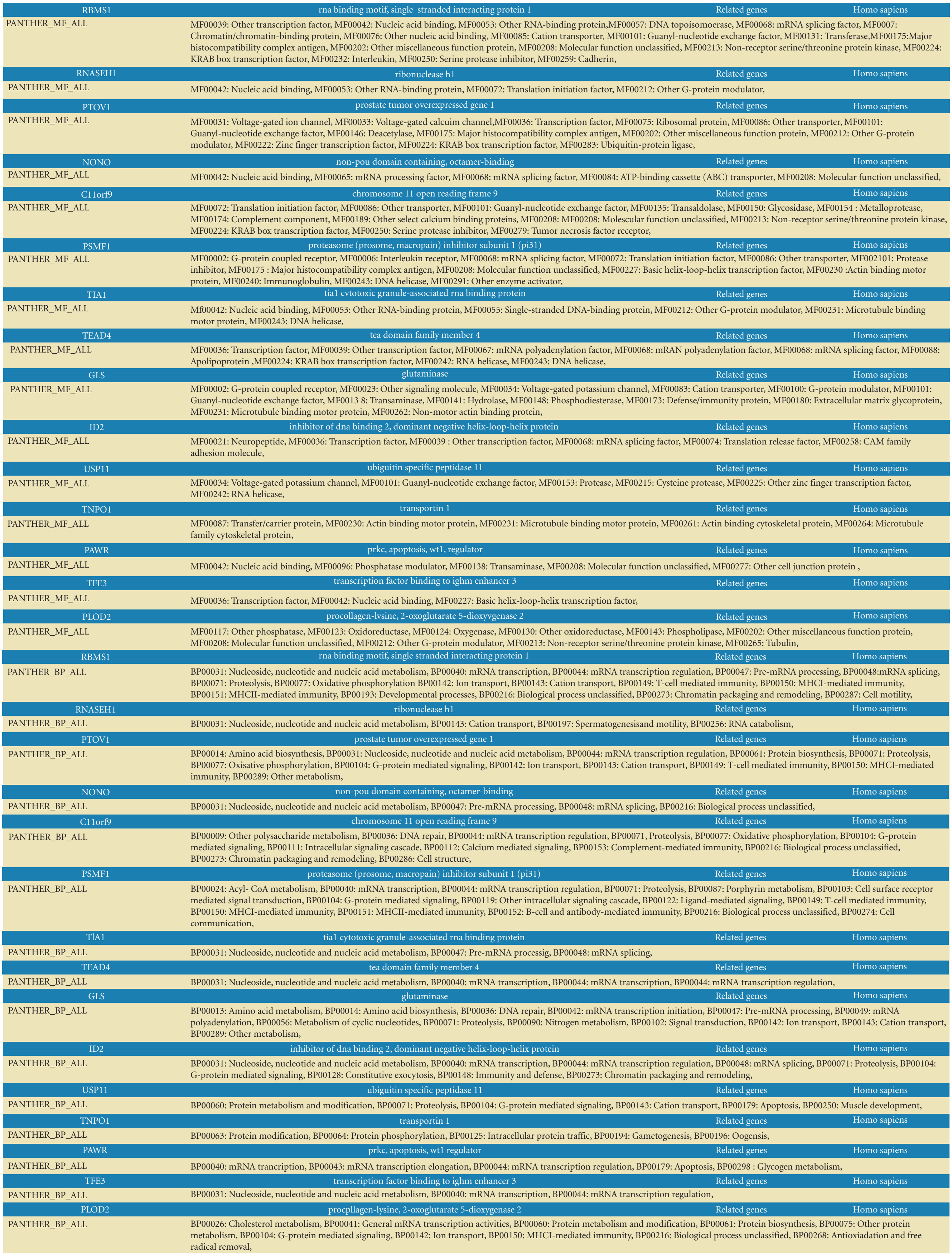
3.3. Constructing Feedback Network of the Distinguished Single Upstream and Downstream Gene. We took into account the feedback relationship and setup ATF2 feedback network, as shown in Figure 3. ATF2 target genes appeared in ATF2 inhibition to CDR2, GLS, and USP11, consistently, its upstream genes also appeared in CDR2, GLS, and USP11 inhibition to ATF2. CDR2 is also called CDR62, where CDR means cerebellar degeneration-related. On Western blot analysis of Purkinje cells and tumor tissue, the anti-Yo sera react with at least 2 antigens, a major species of 62 kD called CDR62 and a minor species of 34 kD called CDR34 [14]. Sahai (1983) demonstrated phosphate-activated glutaminase (GLS) in human platelets [15]. It is the major enzyme yielding glutamate from glutamine. Significance of the enzyme derives from its possible implication in behavior disturbances in which glutamate acts as a neurotransmitter [16]. USP11 is also called UHX1. Swanson et al. (1996) cited evidence indicating that ubiquitin hydrolases play a role in oncogenesis (oncogenes and tumor suppressor gene products are degraded in ubiquitin-dependent pathways) [17]. The relationship of ATF2 with CDR2, GLS, and USP11 represents a negative feedback loop.

3.4. Functional Module Construction of the Distinguished Single Gene. According to ATF2 upstream network, we did DAVID analysis of function cluster, respectively. The DAVID functional annotation clustering results appeared that one ATF2 regulation network was identified as consisting of the ATF2 upstream genes including RBMS1, RNASEH1, PTOV1, NONO, C11orf9, PSMF1, TIA1, TEAD4, GLS, ID2, USP11, TNPO1, PAWR, PLOD2, and TFE3, as shown in Figure 4.

According to Figure 2, it appeared that RBMS1, NONO, PSMF1, TIA1, ID2, PLOD2, TFE3 activate ATF2; whereas RNASEH1, PTOV1, C11orf9, TEAD4, GLS, USP11, TNPO1, and PAWR inhibit ATF2.

RBMS1, NONO, TIA1, ID2, and TFE3 enhance nucleoside, nucleotide, and nucleic acid metabolism because RBMS1, NONO, TIA1, ID2, and TFE3 are involved in these metabolism; PSMF1 activation to ATF2 means the increase of Acyl-CoA metabolism and porphyrin metabolism; PLOD2 activation to ATF2 indicates the progress of cholesterol metabolism and other protein metabolism, as shown in Figure 5.

RNASEH1, PTOV1, and TEAD4 inhibition to ATF2 decreases nucleoside, nucleotide, and nucleic acid metabolism mediated by the three genes; C11orf9 inhibition to ATF2 means the decline of polysaccharide metabolism, whereas GLS represents the weakness of amino acid and cyclic nucleotides metabolism; USP11 inhibition to ATF2 indicates the fall-off in protein metabolism and modification, whereas PAWR in glycogen metabolism, as shown in Figure 5.



RBMS1	rna binding motif, single stranded interacting protein 1	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00039: Other transcription factor, MF00042: Nucleic acid binding, MF00053: Other RNA-binding protein,MF00057: DNA topoisomoerase, MF00068: mRNA splicing factor, MF0007: Chromatin/chromatin-binding protein, MF00076: Other nucleic acid binding, MF00085: Cation transporter, MF00101: Guanyl-nucleotide exchange factor, MF00131: Transferase,MF00175:Major histocompatibility complex antigen, MF00202: Other miscellaneous function protein, MF00208: Molecular function unclassified, MF00213: Non-receptor serine/threonine protein kinase, MF00224: KRAB box transcription factor, MF00232: Interleukin, MF00250: Serine protease inhibitor, MF00259: Cadherin,		
RNASEH1	ribonuclease h1	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00042: Nucleic acid binding, MF00053: Other RNA-binding protein, MF00072: Translation initiation factor, MF00212: Other G-protein modulator,		
PTOV1	prostate tumor overexpressed gene 1	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00031: Voltage-gated ion channel, MF00033: Voltage-gated calcuim channel,MF00036: Transcription factor, MF00075: Ribosomal protein, MF00086: Other transporter, MF00101: Guanyl-nucleotide exchange factor, MF00146: Deacetylase, MF00175: Major histocompatibility complex antigen, MF00202: Other miscellaneous function protein, MF00212: Other G-protein modulator, MF00222: Zinc finger transcription factor, MF00224: KRAB box transcription factor, MF00283: Ubiquitin-protein ligase,		
NONO	non-pou domain containing, octamer-binding	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00042: Nucleic acid binding, MF00065: mRNA processing factor, MF00068: mRNA splicing factor, MF00084: ATP-binding cassette (ABC) transporter, MF00208: Molecular function unclassified,		
C11orf9	chromosome 11 open reading frame 9	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00072: Translation initiation factor, MF00086: Other transporter, MF00101: Guanyl-nucleotide exchange factor, MF00135: Transaldolase, MF00150: Glycosidase, MF00154 : Metalloprotease, MF00174: Complement component, MF00189: Other select calcium binding proteins, MF00208: MF00208: Molescular function unclassified, MF00213: Non-receptor serine/threonine protein kinase, MF00224: KRAB box transcription factor, MF00250: Serine protease inhibitor, MF00279: Tumor necrosis factor receptor,		
PSMF1	proteasome (prosome, macropain) inhibitor subunit 1 (pi31)	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00002: G-protein coupled receptor, MF00006: Interleukin receptor, MF00068: mRNA splicing factor, MF00072: Translation initiation factor, MF00086: Other transporter, MF002101: Protease inhibitor, MF00175 : Major histocompatibility complex antigen, MF00208: Molecular function unclassified, MF00227: Basic helix-loop-helix transcription factor, MF00230 :Actin binding motor protein, MF00240: Immunoglobulin, MF00243: DNA helicase, MF00291: Other enzyme activator,		
TIA1	tia1 cvtotoxic granule-associated rna binding protein	Related genes	Homo sapiens
PANTHER_MF_ALL	Mf00042: Nucleic acid binding, MF00053: Other RNA-binding protein, MF00055: Single-stranded DNA-binding protein, MF00212: Other G-protein modulator, MF00231: Microtubule binding motor protein, MF00243: DNA helicase,		
TEAD4	tea domain family member 4	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00036: Transcription factor, MF00039: Other transcription factor, MF00067: mRNA polyadenylation factor, MF00068: mRAN polyadenylation factor, MF00068: mRNA splicing factor, MF00088: Apolipoprotein ,MF00224: KRAB box transcription factor, MF00242: RNA helicase, MF00243: DNA helicase,		
GLS	glutaminase	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00002: G-protein coupled receptor, MF00023: Other signaling molecule, MF00034: Voltage-gated potassium channel, MF00083: Cation transporter, MF00100: G-protein modulator, MF00101: Guanyl-nucleotide exchange factor, MF0013 8: Transaminase, MF00141: Hydrolase, MF00148: Phosphodiesterase, MF00173: Defense/immunity protein, MF00180: Extracellular matrix glycoprotein, MF00231: Microtubule binding motor protein, MF00262: Non-motor actin binding protein,		
ID2	inhibitor of dna binding 2, dominant negative helix-loop-helix protein	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00021: Neuropeptide, MF00036: Transcription factor, MF00039 : Other transcription factor, MF00068: mRNA splicing factor, MF00074: Translation release factor, MF00258: CAM family adhesion molecule,		
USP11	ubiguitin specific peptidase 11	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00034: Voltage-gated potassium channel, MF00101: Guanyl-nucleotide exchange factor, MF00153: Protease, MF00215: Cysteine protease, MF00225: Other zinc finger transcription factor, MF00242: RNA helicase,		
TNPO1	transportin 1	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00087: Transfer/carrier protein, MF00230: Actin binding motor protein, MF00231: Microtubule binding motor protein, MF00261: Actin binding cytoskeletal protein, MF00264: Microtubule family cytoskeletal protein,		
PAWR	prkc, apoptosis, wt1, regulator	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00042: Nucleic acid binding, MF00096: Phosphatase modulator, MF00138: Transaminase, MF00208: Molecular function unclassified, MF00277: Other cell junction protein ,		
TFE3	transcription factor binding to ighm enhancer 3	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00036: Transcription factor, MF00042: Nucleic acid binding, MF00227: Basic helix-loop-helix transcription factor,		
PLOD2	procollagen-lvsine, 2-oxoglutarate 5-dioxyvgenase 2	Related genes	Homo sapiens
PANTHER_MF_ALL	MF00117: Other phosphatase, MF00123: Oxidoreductase, MF00124: Oxygenase, MF00130: Other oxidoreductase, MF00143: Phospholipase, MF00202: Other miscellaneous function protein, MF00208: Molecular function unclassified, MF00212: Other G-protein modulator, MF00213: Non-receptor serine/threonine protein kinase, MF00265: Tubulin,		
RBMS1	rna binding motif, single stranded interacting protein 1	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00031: Nucleoside, nucleotide and nucleic acid metabolism, BP00040: mRNA transcription, BP00044: mRNA transcription regulation, BP00047: Pre-mRNA processing, BP00048:mRNA splicing, BP00071: Proteolysis, BP00077: Oxidative phosphorylation BP00142: Ion transport, BP00143: Cation transport, BP00149: T-cell mediated immunity, BP00150: MHCI-mediated immunity, BP00151: MHCII-mediated immunity, BP00193: Developmental processes, BP00216: Biological process unclassified, BP00273: Chromatin packaging and remodeling, BP00287: Cell motility,		
RNASEH1	ribonuclease h1	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00031: Nucleoside, nucleotide and nucleic acid metabolism, BP00143: Cation transport, BP00197: Spermatogenesisand motility, BP00256: RNA catabolism,		
PTOV1	prostate tumor overexpressed gene 1	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00014: Amino acid biosynthesis, BP00031: Nucleoside, nucleotide and nucleic acid metabolism, BP00044: mRNA transcription regulation, BP00061: Protein biosynthesis, BP00071: Proteolysis, BP00077: Oxisative phosphorylation, BP00104: G-protein mediated signaling, BP00142: Ion transport, BP00143: Cation transport, BP00149: T-cell mediated immunity, BP00150: MHCI-mediated immunity, BP00289: Other metabolism,		
NONO	non-pou domain containing, octamer-binding	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00031: Nucleoside, nucleotide and nucleic acid metabolism, BP00047: Pre-mRNA processing, BP00048: mRNA splicing, BP00216: Biological process unclassified,		
C11orf9	chromosome 11 open reading frame 9	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00009: Other polysaccharide metabolism, BP00036: DNA repair, BP00044: mRNA transcription regulation, BP00071, Proteolysis, BP00077: Oxidative phosphorylation, BP00104: G-protein mediated signaling, BP00111: Intracellular signaling cascade, BP00112: Calcium mediated signaling, BP00153: Complement-mediated immunity, BP00216: Biological process unclassified, BP00273: Chromatin packaging and remodeling, BP00286: Cell structure,		
PSMF1	proteasome (prosome, macropain) inhibitor subunit 1 (pi31)	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00024: Acyl- CoA metabolism, BP00040: mRNA transcription, BP00044: mRNA transcription regulation, BP00071: Proteolysis, BP00087: Porphyrin metabolism, BP00103: Cell surface receptor mediated signal transduction, BP00104: G-protein mediated signaling, BP00119: Other intracellular signaling cascade, BP00122: Ligand-mediated signaling, BP00149: T-cell mediated immunity, BP00150: MHCI-mediated immunity, BP00151: MHCII-mediated immunity, BP00152: B-cell and antibody-mediated immunity, BP00216: Biological process unclassified, BP00274: Cell communication,		
TIA1	tia1 cytotoxic granule-associated rna binding protein	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00031: Nucleoside, nucleotide and nucleic acid metabolism, BP00047: Pre-mRNA processig, BP00048: mRNA splicing,		
TEAD4	tea domain family member 4	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00031: Nucleoside, nucleotide and nucleic acid metabolism, BP00040: mRNA transcription, BP00044: mRNA transcription, BP00044: mRNA transcription regulation,		
GLS	glutaminase	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00013: Amino acid metabolism, BP00014: Amino acid biosynthesis, BP00036: DNA repair, BP00042: mRNA transcription initiation, BP00047: Pre-mRNA processing, BP00049: mRNA polyadenylation, BP00056: Metabolism of cyclic nucleotides, BP00071: Proteolysis, BP00090: Nitrogen metabolism, BP00102: Signal transduction, BP00142: Ion transport, BP00143: Cation transport, BP00289: Other metabolism,		
ID2	inhibitor of dna binding 2, dominant negative helix-loop-helix protein	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00031: Nucleoside, nucleotide and nucleic acid metabolism, BP00040: mRNA transcription, BP00044: mRNA transcription regulation, BP00048: mRNA splicing, BP00071: Proteolysis, BP00104: G-protein mediated signaling, BP00128: Constitutive exocytosis, BP00148: Immunity and defense, BP00273: Chromatin packaging and remodeling,		
USP11	ubiguitin specific peptidase 11	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00060: Protein metabolism and modification, BP00071: Proteolysis, BP00104: G-protein mediated signaling, BP00143: Cation transport, BP00179: Apoptosis, BP00250: Muscle development,		
TNPO1	transportin 1	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00063: Protein modification, BP00064: Protein phosphorylation, BP00125: Intracellular protein traffic, BP00194: Gametogenesis, BP00196: Oogensis,		
PAWR	prkc, apoptosis, wt1 regulator	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00040: mRNA trancription, BP00043: mRNA transcription elongation, BP00044: mRNA transcription regulation, BP00179: Apoptosis, BP00298 : Glycogen metabolism,		
TFE3	transcription factor binding to ighm enhancer 3	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00031: Nucleoside, nucleotide and nucleic acid metabolism, BP00040: mRNA transcription, BP00044: mRNA transcription regulation,		
PLOD2	procpllagen-lysine, 2-oxoglutarate 5-dioxygenase 2	Related genes	Homo sapiens
PANTHER_BP_ALL	BP00026: Cholesterol metabolism, BP00041: General mRNA transcription activities, BP00060: Protein metabolism and modification, BP00061: Protein biosynthesis, BP00075: Other protein metabolism, BP00104: G-protein mediated signaling, BP00142: Ion transport, BP00150: MHCI-mediated immunity, BP00216: Biological process unclassified, BP00268: Antioxiadation and free radical removal,		

Figure 5: Molecular function and biological process from DAVID.

4. Conclusions

Our method concentrates on and constructs the distinguished single gene network integrated with function prediction analysis by DAVID. For the distinguished single molecular network, we did (1) control and experiment comparison, (2) identification of activation and inhibition networks, (3) construction of upstream and downstream feedback networks, and (4) functional module construction. We tested this method to identify ATF2 regulation network module using data of 45 samples from one and the same GEO dataset. The results demonstrate the effectiveness of such integrated way in terms of developing novel prognostic markers and therapeutic targets.

Acknowledgments

This work was supported by the National Natural Science Foundation in China (no. 60673109 and no. 60871100) and the Teaching and Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, State Key Lab of Pattern Recognition Open Foundation.

References

[1] Y. Wang, T. Joshi, X.-S. Zhang, D. Xu, and L. Chen, "Inferring gene regulatory networks from multiple microarray datasets," *Bioinformatics*, vol. 22, no. 19, pp. 2413–2420, 2006.

[2] D. W. Huang, B. T. Sherman, Q. Tan, et al., "The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists," *Genome Biology*, vol. 8, no. 9, article R183, pp. 1–16, 2007.

[3] D. W. Huang, B. T. Sherman, Q. Tan, et al., "DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists," *Nucleic Acids Research*, vol. 35, web server issue, pp. W169–W175, 2007.

[4] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[5] T. Byrt, J. Bishop, and J. B. Carlin, "Bias, prevalence and kappa," *Journal of Clinical Epidemiology*, vol. 46, no. 5, pp. 423–429, 1993.

[6] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[7] M. B. Humphrey, H. Herrera-Sosa, G. Gonzalez, R. Lee, and J. Bryan, "Cloning of cDNAs encoding human caldesmons," *Gene*, vol. 112, no. 2, pp. 197–204, 1992.

[8] J. A. Williamson, J. M. Bosher, A. Skinner, D. Sheer, T. Williams, and H. C. Hurst, "Chromosomal mapping of the human and mouse homologues of two new members of the AP-2 family of transcription factors," *Genomics*, vol. 35, no. 1, pp. 262–264, 1996.

[9] J. M. Bosher, T. Williams, and H. C. Hurst, "The developmentally regulated transcription factor AP-2 is involved in c-erbB-2 overexpression in human mammary carcinoma," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 3, pp. 744–747, 1995.

[10] P. S. Henthron, C. C. Stewart, T. Kadesch, and J. M. Puck, "The gene encoding human TFE3, a transcription factor that binds the immunoglobulin heavy-chain enhancer, maps to Xp11.22," *Genomics*, vol. 11, no. 2, pp. 374–378, 1991.

[11] Y. Nakagawa, H. Shimano, T. Yoshikawa, et al., "TFE3 transcriptionally activates hepatic IRS-2, participates in insulin signaling and ameliorates diabetes," *Nature Medicine*, vol. 12, no. 1, pp. 107–113, 2006.

[12] S. Parisi, M. J. McKay, M. Molnar, et al., "Rec8p, a meiotic recombination and sister chromatid cohesion phosphoprotein of the Rad21p family conserved from fission yeast to humans," *Molecular and Cellular Biology*, vol. 19, no. 5, pp. 3515–3528, 1999.

[13] G. A. Brar, B. M. Kiburz, Y. Zhang, J.-E. Kim, F. White, and A. Amon, "Rec8 phosphorylation and recombination promote the step-wise loss of cohesins in meiosis," *Nature*, vol. 441, no. 7092, pp. 532–536, 2006.

[14] H. Fathallah-Shaykh, S. Wolf, E. Wong, J. B. Posner, and H. M. Furneaux, "Cloning of a leucine-zipper protein recognized by the sera of patients with antibody-associated paraneoplastic cerebellar degeneration," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 8, pp. 3451–3454, 1991.

[15] S. Sahai, "Glutaminase in human platelets," *Clinica Chimica Acta*, vol. 127, no. 2, pp. 197–203, 1983.

[16] S. B. Prusiner, "Disorders of glutamate metabolism and neurological dysfunction," *Annual Review of Medicine*, vol. 32, pp. 521–542, 1981.

[17] D. A. Swanson, C. L. Freund, L. Ploder, R. R. McInnes, and D. Valle, "A ubiquitin C-terminal hydrolase gene on the proximal short arm of the X chromosome: implications for X-linked retinal disorders," *Human Molecular Genetics*, vol. 5, no. 4, pp. 533–538, 1996.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 648987, 10 pages
doi:10.1155/2009/648987

Methodology Report

Integrating Diverse Information to Gain More Insight into Microarray Analysis

Raja Loganantharaj[1] and Jun Chung[2]

[1] *Bioinformatics Research Lab, University of Louisiana, Lafayette, LA 70504, USA*
[2] *Department of Biochemistry and Molecular Biology, Louisiana State University Health Sciences Center-Shreveport, Shreveport, LA 71130, USA*

Correspondence should be addressed to Raja Loganantharaj, logan@cacs.louisiana.edu and Jun Chung, jchung@lsuhsc.edu

Received 12 December 2008; Revised 23 June 2009; Accepted 17 July 2009

Recommended by Zhenqiu Liu

Microarray technology provides an opportunity to view transcriptions at genomic level under different conditions controlled by an experiment. From an array experiment using a human cancer cell line that is engineered to differ in expression of tumor antigen, integrin $\alpha6\beta4$, few hundreds of differentially expressed genes are selected and are clustered using one of several standard algorithms. The set of genes in a cluster is expected to have similar expression patterns and are most likely to be coregulated and thereby expected to have similar function. The highly expressed set of upregulated genes become candidates for further evaluation as potential biomarkers. Besides these benefits, microarray experiment by itself does not help us to understand or discover potential pathways or to identify important set of genes for potential drug targets. In this paper we discuss about integrating protein-to-protein interaction information, pathway information with array expression data set to identify a set of "important" genes, and potential signal transduction networks that help to target and reverse the oncogenic phenotype induced by tumor antigen such as integrin $\alpha6\beta4$. We will illustrate the proposed method with our recent microarray experiment conducted for identifying transcriptional targets of integrin $\alpha6\beta4$ for cancer progression.

1. Introduction

A micro-array experiment is conducted to study expression profiles of genes in a specimen under different experimental conditions, or over several different time periods. It serves many purposes that include (1) developing a predictive computational model which can be used to predict biomarkers and targets for cancer therapy, (2) gaining some insight on gene regulation when a microarray experiment is conducted in different time points, (3) gaining insight on the genes that may be involved in a situation or disease under investigation, (4) understanding or refining protein-in-protein interaction networks, and (5) annotating uncharacterized genes. In a recent review article on the applications of microarray, Troyanskaya [1] provides some details on the items 2, 4, and 5. Statistical tests are conducted to filter valid signals first and then a subset of genes called differentially expressed genes is selected based on their relative strength or weakness of expression levels with respect to their reference expression values. The differentially expressed probes, which roughly correspond to genes, are reduced to few hundreds while the total number of probes of an experiment is in the order of 20 to 50 thousands.

The set of highly expressed genes are considered to be candidates for biomarkers in a microarray experiment. It is quite difficult to single out the best biomarkers by viewing expression level alone partially due to noise or some association by "guilt." By integrating microarray expression data with other information pertaining to the protein behavior we can improve the quality of decision on biomarkers as has been proposed by Camargo and Azuaje in [2]. Similarly we vcan gain better insight into gene regulation by associating gene expression with protein interaction network with known cancer related pathways.

A significant volume of works has been done that relates or combines microarray data sets and protein-to-protein

Table 1: The high ranking 14 up regulated genes based on the fold changes. For each gene in the list the connectivity in the protein interaction network G is given. None of the ranked upregulated genes are hub nodes.

Genes	Fold changes	Connectivity in G
IL8	5.63	11
S100A3	4.86	4
SOX4	4.54	2
SLCO4A1	4.12	2
MAGEH1	3.77	9
AKR1C1	3.72	2
MAD1L1	3.45	21
IL24	3.35	1
HSPA6	3.25	13
NRCAM	3.18	10
COL6A1	3.07	5
ASPH	3.03	2
TUSC3	2.98	1
PEG10	2.87	1

interaction networks. Based on the expected outcome, these works may be characterized into (1) annotating uncharacterized genes, (2) refining protein-to-protein interaction network, (3) predicting protein to protein interaction, and (4) refining potential biomarkers from array expression. Integrating protein interaction network information with expression data sets along with other information pertaining to a gene has been used [3–7] for annotating uncharacterized protein. In the recent work of Nariai et al. [6], probabilistic approach has been used to integrate protein to protein interaction, array expression, protein motif, gene knockout phenotype data, and protein localization data for predicting the function of an uncharacterized genes.

Microarray expressions data has also been used for refining protein to protein interaction networks. Zhu et al. [8] have used coexpressed genes from microarray data set to filter the neighbors of protein in an interaction network to enhance the degree of functional consensus among the neighbors.

Array expression data sets are used for predicting protein to protein interaction [9, 10]. Recently Soong et al. [10] have used microarray expression to predict protein to protein interaction. A pair of proteins is represented by a feature vector consisting of a concatenation of expression modes or profiles of those proteins along with the Pearson correlation of the expression profiles of these two proteins. They have demonstrated the predictability of using support vector machine with protein to protein interaction of yeast data sets from DIP [11] and 349 yeast microarray expression data sets from GEO [12].

Camargo et al. [12] have integrated array expression data set with expression data for refining potential biomarkers. Their work has some overlapping with our current approach in selecting hub nodes from interaction network and combining with array expression data sets. Their focus, however, was only on refining the biomarkers derived from array expression as opposed to providing insight into potential signal transduction pathways or any other intermediate activities that are not revealed in an array expression.

We take a different approach that compliments the strength of interaction data sets and array expression data sets. The array data sets capture the expression levels at different experimental conditions (or time points) while the information on interaction networks represents experimentally determined and as well as predicted interaction between pairs of proteins in a two-dimensional space without paying attention to the context, the temporal relations, or the process. By bringing two different types of modalities of information together, we believe we can discover some important genes that may have played important roles in the final observation of the array expression.

Suppose we consider a binary case of studying the expression pattern of a cell line of healthy and sick subjects. Examining the differentially expressed genes provides information on which genes are up-or downregulated, and their expression levels. This information alone does not provide insight into deciding interesting set of genes that are either taking part of the progression or the cause of the disease under consideration. We will show how to integrate gene expression with expression patterns with protein to protein interaction, and known genes in disease pathways to gain insight onto a small subset of interesting genes relevant to the disease under investigation.

To illustrate and to apply the idea of integrating microarray data with protein to protein interaction network, and disease related pathways, we use our recent microarray study for identifying transcriptional targets of integrin $\alpha6\beta4$ for cancer progression. Jun Chung and his associates have used the affymetrix HG-U133A_2 to identify transcriptional targets of integrin $\alpha6\beta4$. The goal of the study is to identify $\alpha6\beta4$ transcriptional targets important for breast cancer progression. The $\alpha6\beta4$ integrin, an epithelial-specific integrin, functions as a receptor for the members of the laminin family of extra cellular matrix proteins [13, 14]. While the primary known function of $\alpha6\beta4$ is to contribute to tissue integrity through its ability to mediate the formation of hemidesmosomes (HDs), there is growing evidence suggesting that this integrin also plays a pivotal role in functions associated with cancer progression [13, 14]. For example, high expression of this integrin in women with breast cancer has been shown to correlate significantly with mortality and disease states [13, 14]. However, therapeutic targets of breast cancer that overexpress $\alpha6\beta4$ are not yet well characterized. For this reason, it is essential to elucidate the mechanism by which $\alpha6\beta4$ contributes to breast cancer progression.

We describe the data set, methods, and approaches in Section 2. It is followed by results in Section 3. In Section 4, we summarize and discuss the results.

2. Materials and Methods

2.1. Data. We are focusing on genes of Homo sapiens and their expressions for this experiment. From Affymetrix site

Table 2: The high ranking 14 downregulated genes. For each gene in the list, the connectivity in the protein interaction network G is given. The 5 hub nodes among the ranked down regulated genes are underlined.

Genes	Fold change (inverse)	Connectivity in G
HBE1	9.10	1
H1F0	7.70	7
AZGP1	7.64	3
SNCA	5.24	44
GLUL	5.13	31
TPM1	4.62	17
IGFBP7	4.54	10
MYLK	4.25	28
KCNS3	4.23	1
NGFRAP1	4.12	15
DGKI	3.97	1
IL1RAP	3.92	14
THBS1	3.70	36
MAP1B	3.65	1

at http://www.affymetrix.com/, we have downloaded the annotations (HG-U133A_2.na22.annot) for the genes that are tested in a microarray experiment.

The gene expression data is from our recent microarray experiment using the affymetrix HG-U133A_2 to identify transcriptional targets of integrin $\alpha6\beta4$. Our study here describes the gene expression profile obtained from MDA-MB-435 mock transfectants ($\alpha6\beta4$ negative human cancer cell line) and MDA-MB-435 $\beta4$ integrin transfectants ($\alpha6\beta4$ positive human cancer cell lines). Out of oligonucleotide probe sets representing approximately 22 277 genes, expression of $\beta4$ integrin in MDA-MB-435 cells up regulated 149 genes by twofold or higher. 193 genes are down regulated by over two fold change. We anticipate that microarray data will lead to not only the identification of $\alpha6\beta4$ target genes that are important for breast cancer cell growth, survival, and invasion, but also the discovery of signaling pathways leading to the expression of these genes.

The protein to protein interaction databases include MIPS [15], DIP [11], BIND [16, 17], GRID and I2D [18]. Noise is often a factor in many protein to protein interaction dataset. To minimize the noise and its impacts on the final outcome, we apply ensemble-based method for selecting the interaction. That is, by applying majority voting on interacting pairs from different the database, we can improve the accuracy and minimize the errors in their interaction information. I2D provides experimentally determined and predicted protein to protein interaction with easy to use interface, and thus we have downloaded I2D [18] for homo sapiens genome.

2.2. Data Preprocessing. Suppose we are gathering protein to protein interaction from different sources each with their own accuracy. By combining the results of independent test or source that has prediction accuracy over 50%, we can obtain prediction accuracy better than any one method alone. Suppose we have n independent sources each with some predefined fixed prediction accuracy, say p. Without loss of generality, let us assume n is an odd number. By accepting the decision of majority predictors among n, the combined accuracy is given by the following formula:

$$\text{prediction\_accuracy} = \sum_{i=k}^{i=n} \binom{n}{i} p^{i} (1-p)^{n-i}, \qquad (1)$$

where $k = \lceil (n/2)$.

Suppose nine independent predictors each with prediction accuracy 0.65 are combined by majority votes, the combined prediction accuracy becomes 0.83.

I2D [18] collects and maintains protein to protein interaction from various sources and we have downloaded the interaction information pertaining to Homo Sapiens. By applying the majority votes, we have minimized some plausible noise in the data set.

The microarray experiment was repeated three times and in each repetition the expressions of genes under the following two conditions are measured: (1) integrin negative cell line (control), and (2) integrin positive cell line. Out of the 22 277 genes we have selected only 8512 genes that have valid signal in all measurements. The average of the log ratio between the integrin positive and the control expression in all the repetitions is taken as the expression of a gene. From the expressions, we could create different expression patterns based on the values such as up regulated fold changes over 2 to 3, 3 to 4, and over 4. Among the down regulated genes, we may have the similar groups. For simplicity, we have taken only two patterns, namely, up regulated and down regulated genes. The up regulated genes are those that have fold changes (log of the ratio 2) over 1 and the down regulated are those that have the fold changes (log of the ratio 0.5) less than −1.

2.3. Methods. We have downloaded human protein to protein interaction networks from I2D, which have 13 560 genes that have connectivity from 1 to 694. The connectivity or degree of a node is defined as the number of edges connected to the network and we consider each edge as bidirectional connection. As expected, the interaction follows the scale free distribution. For the purpose of integrating the interaction network with the microarray expression data set, we have extracted a subnetworks from the whole networks that interact with the differentially expressed genes from the experiment. The selected sub networks, which we refer to as G, have 2186 genes including the 190 differentially expressed genes, and 3130 edges. A view of Graph G is shown in Figure 1 as created by Navigator [19]. The up and down regulated genes are shown in red and green, respectively, and the size of each node corresponds to the degree of interaction of that node in the graph.

In a typical microarray analysis, the differentially expressed genes are ranked based on their fold changes and the first few of them as taken as important. We feel that using expression fold change alone to determine the importance of a gene is quite weak. We take a different approach in this paper for discovering a set of important genes under a given

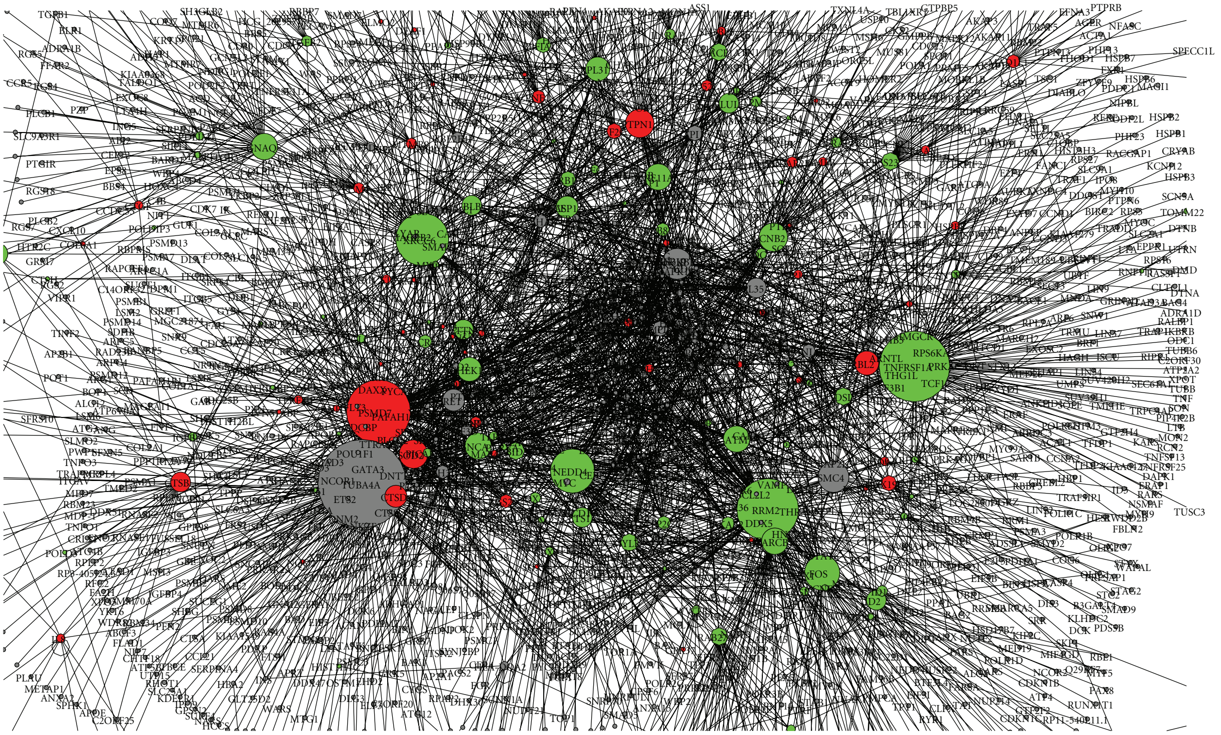


Figure 1: A view of protein to protein interaction associated with the differentially expressed genes. We refer to this graph as G.

experimental condition. We create the subgraphs, say G, of protein to protein interaction networks that is associated with the differentially expressed genes from the microarray experiment. It is generally believed that the connectivity of nodes in G roughly reflects the importance of the gene in the interaction [20]. We found that even the network G has the property of a typical scale free network indicating only a small fraction of the node has large connectivity.

2.3.1. Selecting a Set of Important Genes Based on Topological Structure. In the recent work, Jeong et al. [20] and Twe et al. [21] have suggested that essential proteins are over represented among those proteins having high degree of connectivity, which can be attributed to the central role in mediating interactions among numerous, less connected proteins. Hub nodes in an interaction network are defined as a set of nodes with very high degree of interaction with neighbors and the corresponding threshold for connectivity is defined quite arbitrarily. Vallabhajosyula et al. [22] have studied the issue on selecting hub nodes and the impacts on their functional significance, but unfortunately they were unable to provide and prescriptive definition or method on selecting hub nodes. They, however, stated that the nodes with relatively high degree of interaction are likely to have very high functional significance. In the literature, we found that people have applied varying criteria in selecting the threshold for hub nodes; for example, Batada et al. [23] have defined hub nodes as those connect to over 90% or 95% of the nodes in the network. Biasing from the finding in [22] that the top few percentage of nodes with high degree of interaction has better functional significance, we selected the *hub nodes*; those that are in the top 3% of the nodes ranked based on the decreasing order of connectivity.

We also believe that important genes must also play a role in the stability of the network, that is, removal of such node will break the network into disconnected subnetworks. An *articulation node* in a graph plays the role of connecting or keeping the graph together and the removal of such node separates the graph into subgraphs. Thus the hub genes that play articulation role in an interaction network seem to have more functional significance.

A minimum spanning tree is acyclic graph that connects all the nodes in a network such that the summation of cost in all the edges is minimal and thus eliminates redundant paths among the nodes. A node with high degree of connectivity in minimum spanning tree will indeed play an important role. In a protein interaction network the edge cost is taken to be 1 and we construct a minimal spanning tree using Kruskal's algorithm [24]. We selected the hub nodes from the minimum spanning tree and consider them as important genes too.

As described above, three set of potentially important nodes can be selected from the following different methods: (1) hub nodes from the interaction networks, (2) hub nodes from the set of articulation nodes, and (3) hub nodes from the minimum spanning tree. The nodes satisfying condition 2 are indeed a subset of those satisfying condition 1 and hence we have only two distinct conditions, namely, 2 and 3. We define a set of important genes; those that satisfy either conditions 2 or 3.

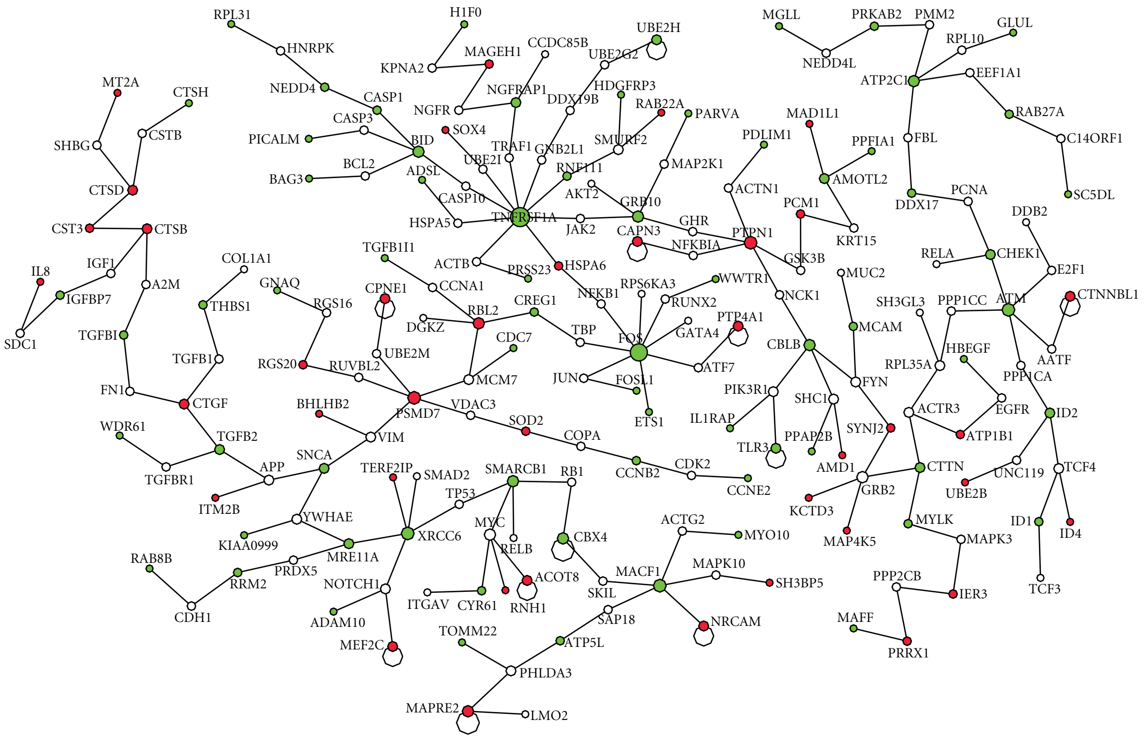


Figure 2: The minimum spanning tree of the network associated with cancer pathway genes. The backbone of the tree is shown. Up-and down regulated genes are shown in red and green color.

2.3.2. Important Genes Based on Pathways and Interaction. Pandey Lab at the Johns Hopkins University and the Institute of Bioinformatics [25] maintains experimentally determined ten cancer signaling pathways for Homo Sapiens, namely, EGFR1, TGF, beta Receptor, TNF, alpha/NF-kB, $\alpha 6\beta 4$ Integrin, ID, Hedgehog, Notch, Wnt, AR, and Kit Receptor. We have obtained the genes in each of the ten cancer pathways and extracted sub network, say G_p, from the interaction network that interacts with any genes in the cancer pathway. The important nodes of G_p include the ones from the three following methods or sources.

(1) Hub nodes of G_p.

(2) Hub nodes of the articulation nodes of G_p.

(3) Hub nodes of the minimum spanning tree created from G_p.

The nodes satisfying condition 2 are indeed a subset of those satisfying condition 1 and hence we have only two distinct conditions, namely, 2 and 3. The important nodes related to cancer pathway are those that satisfy either condition 2 or 3.

Besides examining the important nodes in each graph, we can examine the cliques or near cliques for similar functional association of genes. Han et al. [26] along with many other researchers have used cliques or near cliques in an interaction network to find functional group of genes. A clique is a fully connected subgraph of a graph and find cliques in a network is computationally intractable. For many practical purposes, near cliques are computed.

3. Results

From the microarray experiment, we have two different expression patterns, namely, up-and downregulated genes. The up regulated genes are those that have valid signal across three trials and have expression level over 2 times that of the reference gene. Similarly the down regulated genes are those that have valid signal across three trials and have inverse expression level over 2 with respect to the reference gene. We list the first 14 up and down regulated genes of our experiment in Table 1. We combined the gene expression with gene interaction by selecting subset of the interaction graph that associates with all the differentially expressed genes. The selected subgraphs, which we refer to as G, have 2186 genes including the 190 differentially expressed genes, and 3130 edges. Note that there is no single hub node among the 14 high ranking up regulated nodes of G. On the other hand, there are 5 hub nodes among the high ranking down regulated nodes. There seems to be no correlation among the hub nodes of an interacting graph with highly up or down regulated genes.

From the graph G, we select the set of important genes based on topological structure, which involves selecting the hub nodes and following the procedures described in the previous section. The cutoff connectivity for the hub nodes in G is 16 and there are 60 hub genes out of 2186 genes. Out of the 60 hub nodes, 49 are from the differentially expressed genes (12 of them are up regulated and the rest are down regulated). The graph G has 200 articulation genes and out of which 60 satisfy the hub condition (degree 16 or above). The

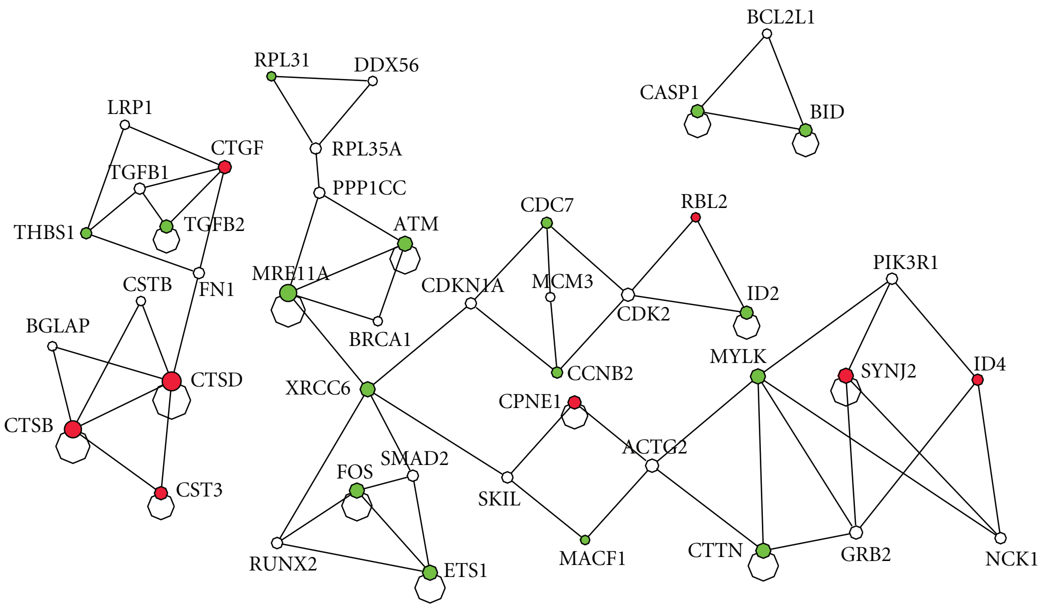


Figure 3: The cliques or near cliques from the cancer pathway related network G_p. The up-and downregulated genes are shown in red and green, respectively.

minimum spanning tree of G was constructed assuming the edge cost is 1. The nodes with connectivity 9 or better in the minimum spanning tree satisfy the hub node property. The minimum spanning tree has 77 hub genes and out of which 17 of them are up regulated and 46 are down regulated. In agreement with conditions 2 and 3 in Section 2, 57 genes are selected as important ones out of which 12 are up regulated and 35 are down regulated. These genes are listed in Table 3.

To discover the important genes related to cancer, we have extracted a sub network, which we call G_p, from G such that each node in G_p is directly associating with any one of the genes in cancer pathways that include EGFR1, TGF, beta Receptor, TNF, alpha/NF-kB, Alpha6 Beta4, Integrin, ID, Hedgehog, Notch, Wnt, AR, and Kit Receptor pathways. The genes in these curated pathways for human are downloaded from their web portal [25]. We found 24 nodes in the network with connectivity 12 or better satisfy the hub node property. The pathway related network G_p has 132 articulation genes out of which 23 are hub genes. The minimum spanning tree of G_p is constructed and the backbone of the minimum spanning tree is shown in Figure 2. The minimum spanning tree has 200 genes and 17 out of these genes have connectivity 4 or better satisfy the hub node property. By combining all these three set of hub genes using ensemble method, we have created the important genes related to pathways and are presented in Table 4.

Besides examining the important genes in G_p, the cancer pathway related network, we searched for cliques or near cliques in the network to examine functionally related genes. The cliques from the network G_p is shown in Figure 3.

Let us examine the interaction among important genes based on topological structure (from Table 3) and between the highly expressed genes from Table 1. The interaction is shown in Figure 4.

The direct interaction among the genes identified as important nodes due to the known cancer pathways is shown in Figure 5.

4. Summary and Discussion

In this paper we have presented a general method for integrating microarray expression with other complementary information related to gene function so that we can understand and infer information about the set of genes that we are interested. Particularly we focused on integrating protein interaction information and pathway related information with microarray expression. We have applied the proposed general methodology to our recent microarray experiment to discover potential drug target that may lead to novel anticancer therapeutics.

Quite a large body of research works is done in integrating expression data with interaction network and other data sets. Many of the works fall into one or some combination of the following categories: (1) annotating uncharacterized genes, (2) refining protein to protein interaction network, (3) predicting protein to protein interaction, and (4) refining potential biomarkers from array expression. The presented work here has some overlaps with the recent work of Camargo et al. [2], which involved in integrating expression data set with expression data set for refining potential biomarkers of array expression and to annotate uncharacterized genes. They have used hub genes of the interaction network to refine biomarkers of the expression data sets.

The interaction network of Homo sapiens is scale free, that is there are few nodes having very high degree of interaction and facilitate other nodes in mediating their functions. Even the subnetwork of the interaction network

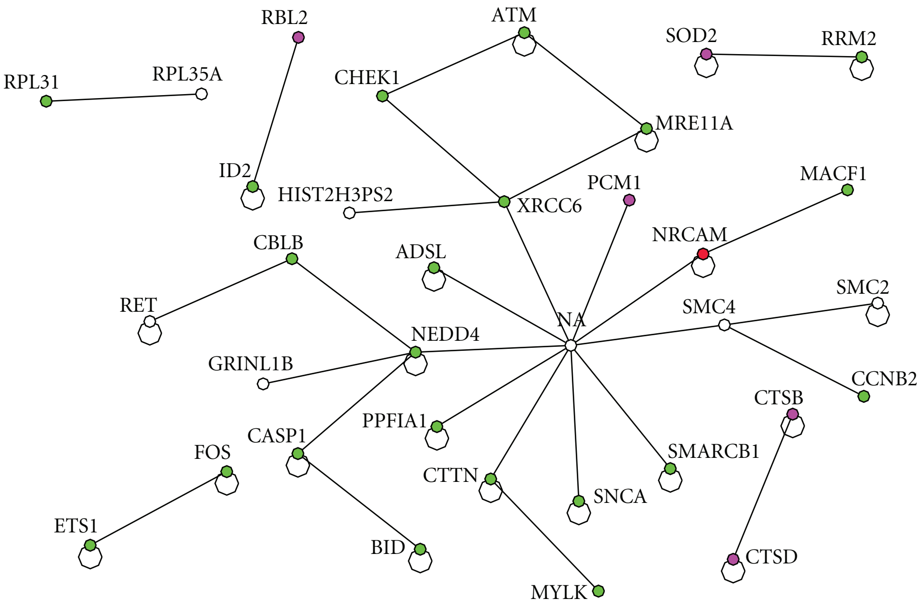


Figure 4: The interaction between the top 14 up regulated genes from Table 1 with the set of important genes based on network topology (Table 3). The red one represents the gene from Table 1. The green colored ones are down regulated and the red and purple ones are up regulated.

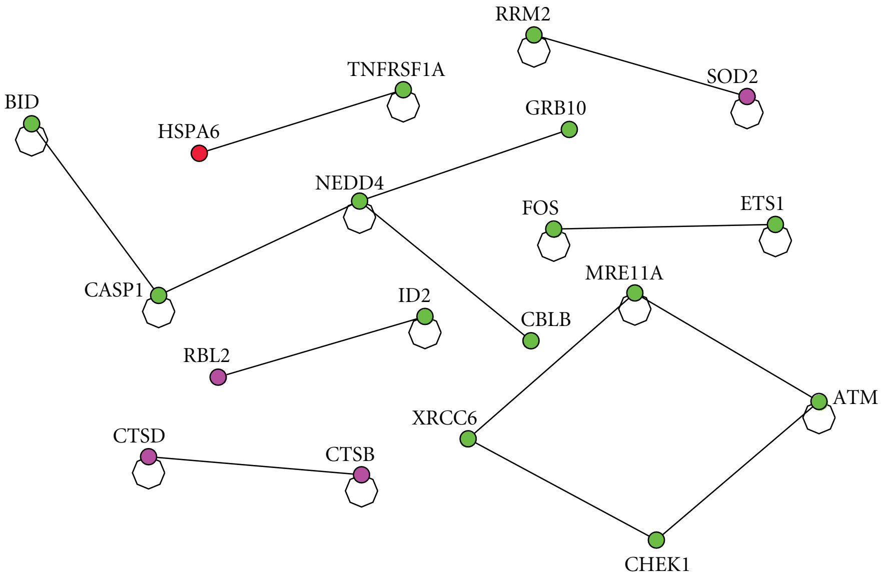


Figure 5: The interaction between the top 14 up regulated genes from Table 1 with the set of important genes based on pathway (Table 4). The red one represents the gene from Table 1. The green colored ones are down regulated and the red and purple ones are up regulated.

that has direct interaction with differentially expressed genes is found to be having the properties of scale free network. Hub nodes in an interaction network are defined as a set of nodes with very high degree of interaction with neighbors and the corresponding threshold for connectivity is defined quite arbitrarily. Biasing from the finding in [22] that the top few percentage of nodes with high degree of interaction has better functional significance, we selected the *hub nodes*; those that are in the top 3% of the nodes ranked based on the decreasing order of connectivity.

From the Homo sapiens interaction network, we have extracted a sub network called G that is associated with the differential expressed genes of our microarray experiment. Hub nodes in an interaction network are important and we selected the first set of hub nodes from G. A set of articulation nodes, which plays the role of stability of the network, is also important. We selected a set of articulation nodes from G. We have constructed a minimum spanning tree from G and we have selected a set of hub nodes from the minimum spanning tree. We created important set of genes based on topological

Table 3: The important set of genes based on topological structure of interaction network. Selecting the nodes that satisfy condition 2 (the articulation nodes among the hub nodes of the network) and condition 3 (the hub nodes of the minimum spanning tree). The inverse of fold changes for down regulated genes is shown. Thus the table includes the genes that are not considered in the experiment or neither up-or downregulated.

Gene	Regulation	Fold change
CPNE1	Up	2.63
CTSB	Up	2.78
CTSD	Up	2.00
MAD1L1	Up	3.45
MEF2C	Up	2.21
PCM1	Up	2.01
PRKAR2B	Up	2.80
PSMD7	Up	2.01
PTPN1	Up	2.06
RBL2	Up	2.62
RGS20	Up	2.20
SOD2	Up	2.10
ADSL	Down	2.23
ATM	Down	2.09
BID	Down	2.09
CASP1	Down	3.07
CBLB	Down	2.32
CCNB2	Down	2.32
CDC7	Down	2.17
CHEK1	Down	2.06
CTTN	Down	2.63
DDX17	Down	2.13
DGCR14	Down	2.04
ETS1	Down	2.50
FOS	Down	2.94
GLUL	Down	5.13
GNAQ	Down	2.39
ID2	Down	2.86
MACF1	Down	2.05
MRE11A	Down	2.61
MYLK	Down	4.25
NEDD4	Down	2.01
PAFAH1B2	Down	2.57
PPFIA1	Down	2.41
PRKAB2	Down	2.56
PRSS23	Down	2.47
RAB27A	Down	2.74
RAB8B	Down	2.27
RPL31	Down	2.08
RRM2	Down	2.32
SMARCB1	Down	2.14
SNCA	Down	5.24
TGFB2	Down	2.06
THBS1	Down	3.70
TNFRSF1A	Down	2.13
TPM1	Down	4.62
XRCC6	Down	2.08
DDX19B	Down	2.09
*GRINL1B	—	—
*HIST2H3PS2	—	—
*NA	—	—
*RET	—	—
*RPL35A	—	—
**SMC2	—	—
**SMC4	—	—
**TPI1	—	—
**TUBA4A	—	—

*These genes are neither up-or downregulated, nor considered in the experiment.

**These genes are from interaction network that satisfy conditions 2 and 3.

Table 4: The important genes of network associated with genes in cancer pathways. These genes are obtained by combining three sets of hub genes from interaction network, articulation nodes, and from the minimum spanning tree of G_p. We show the specific pathway a gene is involved with.

Genes	Regulation	Pathway
CTSB	Up	
CTSD	Up	Tgf_beta,ar
PSMD7	Up	
PTPN1	Up	
RBL2	Up	Tgf_beta
SOD2	Up	Tnf_alpha
ATM	Down	
BID	Down	Tnf_alpha
CASP1	Down	Tnf_beta
CBLB	Down	
CCNB2	Down	Tgf_beta
CHEK1	Down	
ETS1	Down	Tgf_beta,tnf_alpha
FOS	Down	Wnt,ar,kit
GRB10	Down	ar
ID2	Down	Tgf_beta,ar
MRE11A	Down	
NEDD4	Down	Tgf_beta
RRM2	Down	Egfr1
SNCA	Down	
TGFB2	Down	Egfr1,tgf_beta,tnf_alpha,ar
THBS1	Down	Tgf_beta, tnf_alpha, id,wnt
TNFRSF1A	Down	Tgf_beta,notch,kit
XRCC6	Down	

structure of the interaction network. The hub nodes alone in isolation do not reveal any useful information. Similarly the highly ranked up or down regulated genes by themselves do not provide any clue into any potential signaling pathways either.

On the other hand, when we combine the set of important genes based on the interaction topology from Table 3 and the set of highly expressed genes from Table 1, we started to get some insight into potential signal transduction pattern as shown in Figure 4. The highly expressed gene from the experiment NRCAM, neuronal cell adhesion molecule, is directly interacting with another gene NA (neurocanthocytosis) which is recognized as an important gene from the topology and mediating the down regulation of the following set of tumor suppression genes, CHEK1 [27], XRCC6 [28], SMARCB [29], and ATM [30]. The gene NA acts as a hub gene among the set of important genes and it directly interacts with SMARCB and XRCC4, which directly interacts with CHEK1 which in turn directly interacting with ATM. It is notable that down regulation of these tumor suppressor genes by integrin $\alpha6\beta4$ has a significant implication in cancer biology. Poor prognosis has been associated with over expression of integrin $\alpha6\beta4$ and our analysis revealed that loss of these tumor suppressor genes could attribute to malignant phenotype of cancer cells.

Impact of this study lies in the identification and targeting molecular aberrations specific to cancer cells. Many recent studies with targeting a single agent turned out to be a disappointment. This could partly be due to the inability to identify signaling network or loop which is positively or negatively regulated around the single target. To meet this important challenge, a number recent studies are analyzing cancer cell lines and tissue samples to measure alterations at the gene, RNA, and protein level to identify markers and targets for the therapy. While these studies will produce a large amount of data whose analysis is critical in order to understand cancer at the molecular level. For example, a similar microarray analysis of MDA-MB-435 cells that are engineered to differ in integrin $\alpha6\beta4$ expression by Chen et al. leads to the identification of couple of invasion and metastasis related genes such as ENPP2 [31] and S100A4 [32]. What makes our study unique from these works is that we are in a position to identify genes and proteins that are functionally connected to drive malignant properties rather than focusing a single gene because targeting these sub networks will inhibit cancer cell functions important for progression. For example, we found the potentially important $\alpha6\beta4$ target genes associated with cancer pathway as summarized in Table 4. Those genes are associated with TGF-β [33], TNF-α [34], and EGFR1 pathways [35], whose roles in cancer progression have been well established.

In summary, the integration of interaction network with expression of $\alpha6\beta4$ integrin in MDA-MB-435 cancer cells reveals the importance of NRCAM, which we would not have discovered with the expression information alone. Further, the interaction network in Figure 4 helps us to understand how the tumor suppression genes CHEK1, XRCC6, SMARCB, ATM, CHEK1 were down regulated by integrin $\alpha6\beta4$. Finally, we envision the discovery of interaction network triggered from tumor antigen such as integrin $\alpha6\beta4$ will lead to the development of novel anticancer therapeutics by targeting signaling molecules associated with interaction network.

References

[1] O. G. Troyanskaya, "Putting microarrays in a context: integrated analysis of diverse biological data," *Briefings in Bioinformatics*, vol. 6, no. 1, pp. 34–43, 2005.

[2] A. Camargo and F. Azuaje, "Identification of dilated cardiomyopathy signature genes through gene expression and network data integration," *Genomics*, vol. 92, no. 6, pp. 404–413, 2008.

[3] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein, "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 14, pp. 8348–8353, 2003.

[4] S. V. Date and E. M. Marcotte, "Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages," *Nature Biotechnology*, vol. 21, no. 9, pp. 1055–1062, 2003.

[5] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble, "Kernel-based data fusion and its application to protein function prediction in yeast," in *Proceedings of the Pacific Symposium on Biocomputing (PSB '04)*, pp. 300–311, Big Island of Hawaii, Hawaii, USA, January 2004.

[6] N. Nariai, E. D. Kolaczyk, and S. Kasif, "Probabilistic protein function prediction from heterogeneous genome-wide data," *PLoS ONE*, vol. 2, no. 3, article e337, 2007.

[7] H. N. Chua, W.-K. Sung, and L. Wong, "An efficient strategy for extensive integration of diverse biological data for protein function prediction," *Bioinformatics*, vol. 23, no. 24, pp. 3364–3373, 2007.

[8] M. Zhu, L. Gao, Z. Guo, et al., "Globally predicting protein functions based on co-expressed protein-protein interaction networks and ontology taxonomy similarities," *Gene*, vol. 391, no. 1-2, pp. 113–119, 2007.

[9] X. Lin, M. Liu, and X.-W. Chen, "Assessing reliability of protein-protein interactions by integrative analysis of data in model organisms," *BMC Bioinformatics*, vol. 10, article S4, 2009.

[10] T.-T. Soong, K. O. Wrzeszczynski, and B. Rost, "Physical protein-protein interactions predicted from microarrays," *Bioinformatics*, vol. 24, no. 22, pp. 2608–2614, 2008.

[11] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, no. 1, pp. 303–305, 2002.

[12] T. Barrett, D. B. Troup, S. E. Wilhite, et al., "NCBI GEO: archive for high-throughput functional genomic data," *Nucleic Acids Research*, vol. 37, pp. D885–D890, 2009.

[13] A. M. Mercurio, R. E. Bachelder, I. Rabinovitz, K. L. O'Connor, T. Tani, and L. M. Shaw, "The metastatic odyssey: the integrin connection," *Surgical Oncology Clinics of North America*, vol. 10, no. 2, pp. 313–328, 2001.

[14] E. A. Lipscomb and A. M. Mercurio, "Mobilization and activation of a signaling competent $\alpha6\beta4$ integrin underlies its

contribution to carcinoma progression," *Cancer and Metastasis Reviews*, vol. 24, no. 3, pp. 413–423, 2005.

[15] P. Pagel, S. Kovac, M. Oesterheld, et al., "The MIPS mammalian protein-protein interaction database," *Bioinformatics*, vol. 21, no. 6, pp. 832–834, 2005.

[16] G. D. Bader, D. Betel, and C. W. V. Hogue, "BIND: the biomolecular interaction network database," *Nucleic Acids Research*, vol. 31, no. 1, pp. 248–250, 2003.

[17] R. C. Willis and C. W. Hogue, "Searching, viewing, and visualizing data in the Biomolecular Interaction Network Database (BIND)," *Current Protocols in Bioinformatics*, chapter 8: unit 8.9, 2006.

[18] K. R. Brown and I. Jurisica, "Unequal evolutionary conservation of human protein interactions in interologous networks," *Genome Biology*, vol. 8, no. 5, article R95, 2007.

[19] "NAViGaTOR 2.0," 2008, http://ophid.utoronto.ca/navigator/index.html.

[20] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[21] K. L. Tew, X. L. Li, and S. H. Tan, "Functional centrality: detecting lethality of proteins in protein interaction networks," *Genome Informatics*, vol. 19, pp. 166–177, 2007.

[22] R. R. Vallabhajosyula, D. Chakravarti, S. Lutfeali, A. Ray, and A. Raval, "Identifying hubs in protein interaction networks," *PLoS ONE*, vol. 4, no. 4, article e5344, 2009.

[23] N. N. Batada, T. Reguly, A. Breitkreutz, et al., "Stratus not altocumulus: a new view of the yeast protein interaction network," *PLoS Biology*, vol. 4, no. 10, p. e317, 2006.

[24] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, 2nd edition, 2001.

[25] "NetPath," http://www.netpath.org.

[26] K. Han, G. Cui, and Y. Chen, "Identifying functional groups by finding cliques and near-cliques in protein interaction networks," in *Proceedings of the Frontiers in the Convergence of Bioscience and Information Technologies (FBIT '07)*, pp. 159–164, 2007.

[27] X. Q. Wang, E. J. Stanbridge, X. Lao, Q. Cai, S. T. Fan, and J. L. Redpath, "p53-dependent Chk1 phosphorylation is required for maintenance of prolonged G2 arrest," *Radiation Research*, vol. 168, no. 6, pp. 706–715, 2007.

[28] P. Willems, K. De Ruyck, R. Van den Broecke, et al., "A polymorphism in the promoter region of Ku70/XRCC6, associated with breast cancer risk and oestrogen exposure," *Journal of Cancer Research and Clinical Oncology*, vol. 135, no. 9, pp. 1159–1168, 2009.

[29] C. W. M. Roberts and J. A. Biegel, "The role of SMARCB1/INI1 in development of rhabdoid tumor," *Cancer Biology and Therapy*, vol. 8, no. 5, pp. 412–416, 2009.

[30] R. T. Abraham, "Cell cycle checkpoint signaling through the ATM and ATR kinases," *Genes and Development*, vol. 15, no. 17, pp. 2177–2196, 2001.

[31] M. Chen and K. L. O'Connor, "Integrin $\alpha 6\beta 4$ promotes expression of autotaxin/ENPP2 autocrine motility factor in breast carcinoma cells," *Oncogene*, vol. 24, no. 32, pp. 5125–5130, 2005.

[32] M. Chen, M. Sinha, B. A. Luxon, A. R. Bresnick, and K. L. O'Connor, "Integrin $\alpha 6\beta 4$ controls the expression of genes associated with cell motility, invasion, and metastasis, including S100A4/metastasin," *The Journal of Biological Chemistry*, vol. 284, no. 3, pp. 1484–1494, 2009.

[33] G. J. Prud'homme, "Pathobiology of transforming growth factor β in cancer, fibrosis and immunologic disease, and therapeutic considerations," *Laboratory Investigation*, vol. 87, no. 11, pp. 1077–1091, 2007.

[34] I. Zidi, S. Mestiri, A. Bartegi, and N. B. Amor, "TNF-α and its inhibitors in cancer," *Medical Oncology*, pp. 1–14, 2009.

[35] L. Kopper, "Lapatinib: a sword with two edges," *Pathology and Oncology Research*, vol. 14, no. 1, pp. 1–8, 2008.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 632786, 7 pages
doi:10.1155/2009/632786

Research Article

Developing Prognostic Systems of Cancer Patients by Ensemble Clustering

Dechang Chen,[1] Kai Xing,[2] Donald Henson,[3] Li Sheng,[4] Arnold M. Schwartz,[5] and Xiuzhen Cheng[2]

[1] *Division of Epidemiology and Biostatistics, Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA*
[2] *Department of Computer Science, The George Washington University, Washington DC 20052, USA*
[3] *The George Washington University Cancer Institute, The George Washington University, Washington DC 20037, USA*
[4] *Department of Mathematics, Drexel University, Philadelphia, PA 19104, USA*
[5] *Department of Pathology, The George Washington University Medical Center, Washington DC 20037, USA*

Correspondence should be addressed to Dechang Chen, dchen@usuhs.mil

Received 7 January 2009; Accepted 27 March 2009

Recommended by Zhenqiu Liu

Accurate prediction of survival rates of cancer patients is often key to stratify patients for prognosis and treatment. Survival prediction is often accomplished by the TNM system that involves only three factors: tumor extent, lymph node involvement, and metastasis. This prediction from the TNM has been limited, because other potential prognostic factors are not used in the system. Based on availability of large cancer datasets, it is possible to establish powerful prediction systems by using machine learning procedures and statistical methods. In this paper, we present an ensemble clustering-based approach to develop prognostic systems of cancer patients. Our method starts with grouping combinations that are formed using levels of factors recorded in the data. The dissimilarity measure between combinations is obtained through a sequence of data partitions produced by multiple use of PAM algorithm. This dissimilarity measure is then used with a hierarchical clustering method in order to find clusters of combinations. Prediction of survival is made simply by using the survival function derived from each cluster. Our approach admits multiple factors and provides a practical and useful tool in outcome prediction of cancer patients. A demonstration of use of the proposed method is given for lung cancer patients.

1. Introduction

Accurate prediction of outcomes or survival rates of cancer patients is often key to stratify patients for prognosis and treatment. Outcomes of patients are usually generated using standard survival functions and various factors recorded in the database (such as SEER [1] or NCDB [2]) that have prognostic potential. All prognostic factors become integrated through determination of the outcome according to the survival rate. This integration leads to a prognostic system that can be used to predict outcomes of any new patients. Clearly, a crucial question is how can one form a powerful prognostic system for cancer patients? The traditional answer to this question is to use the TNM system [3] that involves only three factors: tumor extent, lymph node involvement, and metastasis. However, the outcome prediction from the TNM has been limited, mainly because any other potential prognostic factors are not used in the system.

In this paper, we propose a computer-based prognostic system for cancer patients that admit multiple prognostic factors. Here is idea of our approach: (i) we partition patients from a cancer dataset into "natural" groups such that patients in the same group are more similar in survival than patients from different groups; (ii) once "natural" groups are obtained, a survival function for each group can be estimated by a standard method. Our prognostic system then consists of groups of patients and survival functions associated with the groups.

The first step (i) is the key to the entire process. Mathematically, this step is equivalent to performing a cluster analysis on a cancer dataset. However, this type of cluster analysis is different from traditional clustering approaches, which may be elaborated below. Suppose, after some simple management, a typical record for a patient contained in a cancer dataset is of the form: $X, X_1, \ldots, X_m$, where X is the recorded patient's survival time, which can be a censored time, and $X_1, \ldots, X_m$ are measurements made on m risk factors or variables such as tumor size, gender, and age. Cluster analysis rising in (i) means that clusters of patients are sought such that patients in the same cluster are more similar in their lifetime T than patients from different groups. Here the connection between T and the observed time X is described as follows: $T = X$ if X is an actual time to death due to the cancer under study; $T > X$ otherwise (in this case X is a censored time). Therefore, cluster analysis from (i) is not equivalent to partitioning the set of vectors $\{(X, X_1, \ldots, X_k)\}$ or the set $\{(X_1, \ldots, X_k)\}$ which could be suggested by traditional clustering methods.

The above discussed difference between the cluster analysis in (i) and the traditional clustering indicates that clustering required in (i) may not be a trivial task. Other potential challenges in accomplishing (i) include presence of a high percentage of censored observations, different types of risk factors or variables, and a large dataset size [4–6]. For example, an SEER dataset of lung cancer patients diagnosed from 1973 through 2002 has more than 500 000 patients, comprises more than 30% records with censored survival times, and involves more than 80 variables that are either on the continuous, or ordinal, or nominal scale.

To overcome the above mentioned possible difficulties, we consider subsets of a cancer data, based on combinations of levels of some known key factors. This reduces the complexity in establishing prognostic systems. We then group these subsets by a hierarchical clustering algorithm, where the distance measure between two subsets is learnt through multiple clustering based on Partitioning Around Medoids (PAM) of Kaufman and Rousseeuw [7].

The rest of the paper is organized as follows. In Section 2, we briefly review some necessary elements of clustering and survival analysis. In Section 3, we present our algorithm of clustering of cancer data. An application of our algorithm to establishing a prognostic system for lung cancer patients is provided in Section 4. And finally our conclusion is given in Section 5.

2. Some Elements of Clustering and Survival Analysis

Clustering may be viewed as a process of finding natural groupings of objects. Commonly used clustering procedures fall into two categories: partitioning approaches and hierarchical approaches. A partitioning approach assigns objects into a group or cluster through optimizing some criterion. A hierarchical approach produces a hierarchy of groups or clusters. In this paper, we will use the PAM algorithm (a partitioning algorithm) and linkage methods (special cases of Hierarchical clustering techniques). They will be briefly reviewed in this section. Also reviewed in this section are some notations of censoring and survival functions. Censored survival times often occur in a cancer dataset and represent on type of incomplete data. A survival function provides a probability of survival to certain times for a cancer patient.

2.1. PAM. Partitioning is one of the major clustering approaches. PAM is a partitioning method operating on a dissimilarity matrix, a matrix of pairwise dissimilarities or distances between objects. It starts from selecting initial K (a predetermined number) representative objects, or medoids, assigning each data object to the nearest medoid, and then iteratively replaces one of the medoids by one of the nonmedoids which leads to a reduction in the sum of the distances of the objects to their closest medoids. The similarity measure here includes, as a special case, the Euclidean distance, which is used with the K-means algorithm. PAM is more robust than the K-means approach, because it employs as cluster centers the medoids not the means, and minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances.

2.2. Linkage Methods. Hierarchical clustering procedures are the most commonly used clustering methods in practice. Commonly used linkage methods include single linkage (SL), complete linkage (CL), and average linkage (AL). They are special cases of agglomerative clustering techniques, operate on a given dissimilarity matrix, and follow the same procedure beginning with the individual objects, at each intermediate step two least dissimilar clusters are merged into a single cluster, producing one less cluster at the next higher level [8]. The difference among the linkage methods lies in the dissimilarity measures between two clusters, which are used to merge clusters. SL, CL, and AL define, respectively, the dissimilarity between two clusters to be the minimum distance between objects from these two clusters, the maximum distance between objects from these two clusters, and the average distance between objects in the two clusters. The output of a linkage method is often summarized into a plot where the nested clusters are graphically represented as a tree, called a dendrogram. The branches in the tree represent clusters. Two clusters merge at a height along a dissimilarity axis that is equal to the dissimilarity between the two clusters.

2.3. Censoring. Cancer data are often time-to-event data that present themselves in different ways, imposing great challenges in analysis. One special feature of a large cancer data set is censoring [9]. Censored observations come from the mechanism of monitoring the progress of patients from some point in time, such as the time a surgical procedure is performed or a treatment regimen is initiated, until the occurrence of some predefined event such as death. Censoring comes in many different forms and right censoring is widely used in clinical studies. Right censoring is used to record the amount of time elapsing between the point at which the patient entered the study and the point at which he or she experienced one of the following three events: the event of interest (e.g., death for most of the cancer studies);

loss to follow-up for some reason such as death caused by a health problem other than the one being considered or having moved to another locality; alive at the time the study is terminated. The time elapsing between enrollment in the study and experiencing one of these three events is called the patient's survival time. A survival time is censored if it is not the actual time between enrollment and experiencing the event of interest. Given a censored survival time for a patient, all we know about the lifetime of the patient is that it is greater than some value. Censored survival times provide only a portion of information on the actual lifetimes.

2.4. Survival Function. A patient's lifetime T is a random variable having a probability distribution. In addition to the commonly used probability density function, the distribution of T can also be characterized by the survival function, defined to be $S(t) = P(T > t)$. The function $S(t)$ provides the probability of surviving beyond t. The survival function is usually estimated by a nonparametric method referred to as the Kaplan-Meier estimator [10]. An estimated survival function may be portrayed visually in a survival curve graph. A direct comparison of several survival curves can be conducted by examining the curves appearing in a single graph. A theoretical comparison of several survival functions can be made by conducting a commonly used test such as the log-rank test, Gehan's test [11], Breslow's test [12], and test of Tarone and Ware [13].

3. Algorithm of Clustering of Cancer Data

A key issue related to clustering is how one measures the dissimilarity between objects. Most clustering algorithms presume a measure of dissimilarity. For example, the K-means clustering uses Euclidean distance as a dissimilarity measure. Since cancer data involve censored survival times, a direct use of existing clustering algorithms is not applicable. With cancer data, it is important to find a way to define objects and dissimilarity between objects prior to execution of any clustering algorithm.

Suppose, for a cancer data set, a certain number of factors have been selected for consideration. Various combinations can then be formed by using levels of factors. Specifically, a combination is a subset of the data that correspond to one level of each factor. Suppose there are available a total of N combinations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. A combination plays a role of an object in the cluster analysis. When developing a prognostic system, we need to find groups of patients such that patients within each group are more similar in survival than patients from different groups. Assuming that all patients coming from the same combination have a similar survival rate, then this is equivalent to finding natural groups of combinations.

After objects become available, we can start to define a dissimilarity measure between objects. A dissimilarity measure $\text{dis}(\mathbf{x}_i, \mathbf{x}_j)$ is a nonnegative function that is symmetric with respect to $\mathbf{x}_i$ and $\mathbf{x}_j$. For cancer data, a direct method is to define the dissimilarity between two combinations in light of the difference between the two corresponding survival functions, and the details follow below. Given two combinations $\mathbf{x}_i$ and $\mathbf{x}_j$, testing if there is a difference between the corresponding two survival functions can be done by conducting a commonly used test such as the log-rank test. It is known that a smaller value of a test statistic shows a stronger evidence of no difference. Thus we can define dissimilarity or "distance" between $\mathbf{x}_i$ and $\mathbf{x}_j$ to be

$$\text{dis}_0\left(\mathbf{x}_i, \mathbf{x}_j\right) = \text{ the value of a test statistic.} \qquad (1)$$

Clearly, $\text{dis}_0(\mathbf{x}_i, \mathbf{x}_j) > 0$. This dissimilarity measure in (1) is not the one we actually use when developing cancer predictive systems. In fact, we will use the dissimilarity (1) for the PAM algorithm only and generate a learnt dissimilarity measure for the cancer data through combining assignments from multiple clusterings based on the PAM algorithm. A learnt measure should be more realistic than that in (1). This learnt dissimilarity will then be used with a hierarchical clustering algorithm to produce prognostic systems.

Below we first discuss learning dissimilarity from the use of PAM. And then we present an ensemble clustering algorithm using the learnt dissimilarity and linkage methods to develop prognostic systems for cancer patients.

3.1. Learning Dissimilarity from Data. Different choices of dissimilarity functions can lead to quite different clustering results. Prior knowledge is often helpful in selecting an appropriate dissimilarity measure for a given problem. However, it is possible to learn a dissimilarity function from the data. We describe such a procedure as follows.

Partitioning methods are usually not stable in the sense that the final results often depend on initial assignments. However, if two objects are assigned to the same cluster by a high percentage of the times of use of the same partitioning method, it is then very likely that these two objects come from a common "hidden" group. This heuristic implies that the "actual" dissimilarity between two objects may be derived by combining the various clustering results from repeated use of the same partitioning technique. Here we formalize this combining process using the PAM partitioning method.

For the data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, we can select K initial medoids and then run PAM with the dissimilarity measure (1) to partition the data into K clusters. It is known that the final assignment usually depends on the initial reallocation. Now we run PAM N times. Each time a number K is randomly picked from a given interval $[K_1, K_2]$. By doing this, we may end up with N possibly different final assignments. Given two objects $\mathbf{x}_i$ and $\mathbf{x}_j$, let p_{ij} denote the probability that they are not placed into the same cluster by the final assignment of a run of PAM. This probability p_{ij} can be estimated by using the results of repeated PAM clustering. Define $\delta_l(i, j) = 1$ if the lth use of the PAM algorithm does not assign $\mathbf{x}_i$ and $\mathbf{x}_j$ into the same cluster; and $\delta_l(i, j) = 0$ otherwise. Then $\delta_1(i, j), \delta_2(i, j), \ldots, \delta_N(i, j)$ are i.i.d Bernoulli (p_{ij}). It is well known that the best unbiased estimator of p_{ij} is $\sum_{l=1}^{N} \delta_l(i, j)/N$. This estimate will be used as the dissimilarity measure between $\mathbf{x}_i$ and $\mathbf{x}_j$, that is,

$$\text{dis}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \frac{\sum_{l=1}^{N} \delta_l(i, j)}{N}. \qquad (2)$$

```
(1) Given $N$, $K_1$, and $K_2$, run the PAM clustering method $N$ times with each $K$ randomly chosen from $[K_1, K_2]$.
(2) Construct the pairwise dissimilarity measure $\text{dis}(\mathbf{x}_i, \mathbf{x}_j)$ by using the (2).
(3) Cluster the $n$ objects by applying a linkage method and the dissimilarity measure $\text{dis}(\mathbf{x}_i, \mathbf{x}_j)$ from Step 2.
```

Algorithm 1: Ensemble algorithm of clustering of cancer data.

Table 1: Lung cancer data of 90,214 patients. Survival time is measured in months. Here, adeno, squamous, large, and small represent adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and small cell carcinoma, respectively.

Patient	Survival time (X)	Stage (X_1)	Grade (X_2)	Histology (X_3)	Gender (X_4)
1	64	1	2	squamous	1
2	24	1	3	large	1
3	24	2	3	squamous	1
4	8	1	2	squamous	1
5	16	3	3	squamous	2
6	143	3	2	adeno	2
7	6	3	3	small	2
8	1	4	4	small	1
9	9	1	3	adeno	2
—	—	—	—	—	—
—	—	—	—	—	—
90211	1	1	3	squamous	1
90212	2	1	2	adeno	1
90213	62	2	3	adeno	1
90214	4	4	4	squamous	2

Table 2: A list of 128 combinations based on factor levels. Here, adeno, squamous, large, and small represent adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and small cell carcinoma, respectively.

Group name	Stage (X_1)	Grade (X_2)	Histology (X_3)	Gender (X_4)	Sample size
Comb 1	I	1	adeno	1	1008
Comb 2	I	1	adeno	2	1426
Comb 3	I	1	squamous	1	430
Comb 4	I	1	squamous	2	187
Comb 5	I	1	large	1	8
Comb 6	I	1	large	2	4
Comb 7	I	1	small	1	2
Comb 8	I	1	small	2	2
Comb 9	I	2	adeno	1	2389
Comb 10	I	2	adeno	2	2662
—	—	—	—	—	—
—	—	—	—	—	—
Comb 123	IV	4	squamous	1	163
Comb 124	IV	4	squamous	2	70
Comb 125	IV	4	large	1	1503
Comb 126	IV	4	large	2	911
Comb 127	IV	4	small	1	4246
Comb 128	IV	4	small	2	3368

A smaller value of $\text{dis}(\mathbf{x}_i, \mathbf{x}_j)$ is expected to imply a bigger chance that $\mathbf{x}_i$ and $\mathbf{x}_j$ come from the same "hidden" group.

3.2. Clustering of Cancer Data. With the learnt dissimilarity (2) between the combinations, we can choose a clustering method to form "natural" groups of the combinations. For flexibility and easy interpretation in practice, we choose a hierarchical clustering approach. The final ensemble algorithm of clustering of cancer data (EACCD) is shown in Algorithm 1. Here the word ensemble refers to the sequence of the PAM procedures involved in the method.

Early issues on ensemble clustering were discussed in [14] from the perspective of evidence accumulation. The work in [15] combined the K-means algorithm and linkage methods to form an ensemble method of discovering sample classes using gene expression profiles.

4. Results on Lung Cancer

4.1. Dataset. In this study, we used the SEER data [1] containing records of lung cancer patients diagnosed from the year 1988 through 1998 and examined the following factors: AJCC stage, grade, histological type, and gender. We considered four factors, X_1, X_2, X_3, and X_4 that were set to be stage, grade, histological type, and gender, respectively. For simplicity, we only investigated the following four important levels of X_3: adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and small cell carcinoma. The levels of other three variables were those commonly used in the lung cancer study. Factor X_1 had four levels: I, II, III, and IV; factor X_2 had four levels: 1, 2, 3, and 4; and factor X_4 had two levels: 1 (male) and 2 (female). The final data we actually used involve 90, 214 patients. A portion of the data, in terms of X(survival time), X_1, X_2, X_3, and X_4, is provided in]Table 1.

Before running our algorithm EACCD, we used the levels of four factors X_1, X_2, X_3, and X_4 to partition the dataset into $128(= 4 \times 4 \times 4 \times 2)$ combinations, shown in Table 2. Due to the approximation of the chi-square distribution to the log-rank test statistic, a combination containing less than 100 patients was dropped from our study. In this case, no further analysis was done for these combinations, and our attention was paid to all the other combinations that have a size equal to or larger than 100. For example, Comb 5, Comb 6, Comb 7, Comb 8, Comb 124, as shown in Table 2, were

Table 3: Seven groups produced by cutting the dendrogram in Figure 1 at the height 0.93.

Group	Combinations	Sample size
Group 1	Stage I, Grade 1, adeno Stage I, Grade 2, adeno Stage I, Grade 2, squamous, female Stage I, Grade 3, adeno, female Stage I, Grade 4, adeno, female	11303
Group 2	Stage I, Grade 1, squamous Stage I, Grade 2, squamous, male Stage I, Grade 3, adeno, male Stage I, Grade 3, squamous Stage I, Grade 3, large cells, female Stage I, Grade 4, adeno, male Stage I, Grade 4, large cells Stage II, Grade 1, adeno, female Stage II, Grade 2, adeno, female Stage II, Grade 2, squamous, female	13431
Group 3	Stage I, Grade 1, squamous, male Stage I, Grade 3, large cells, male Stage I, Grade 4, squamous, male Stage II, Grade 1, adeno, male Stage II, Grade 2, adeno, male Stage II, Grade 2, squamous, male Stage II, Grade 3, adeno Stage II, Grade 3, squamous Stage II, Grade 4, large cells	4522
Group 4	Stage I, Grade 4, small cells Stage II, Grade 4, small cells Stage III, Grade 1, adeno Stage III, Grade 2, adeno	4291
Group 5	Stage III, Grade 1, squamous Stage III, Grade 2, squamous Stage III, Grade 3 Stage III, Grade 4, adeno Stage III, Grade 4, squamous, male Stage III, Grade 4, large cells Stage III, Grade 4, small cells	24951
Group 6	Stage IV, Grade 1, adeno, male Stage IV, Grade 1, squamous, male Stage IV, Grade 2, adeno Stage IV, Grade 2, squamous, male Stage IV, Grade 3, adeno, female Stage IV, Grade 3, squamous, female Stage IV, Grade 3, small cells Stage IV, Grade 4, adeno Stage IV, Grade 4, small cells	18215
Group 7	Stage IV, Grade 2, squamous, female Stage IV, Grade 3, adeno, male Stage IV, Grade 3, squamous, male Stage IV, Grade 3, large cells Stage IV, Grade 4, squamous, male Stage IV, Grade 4, large cells	12237

Table 3: Continued.

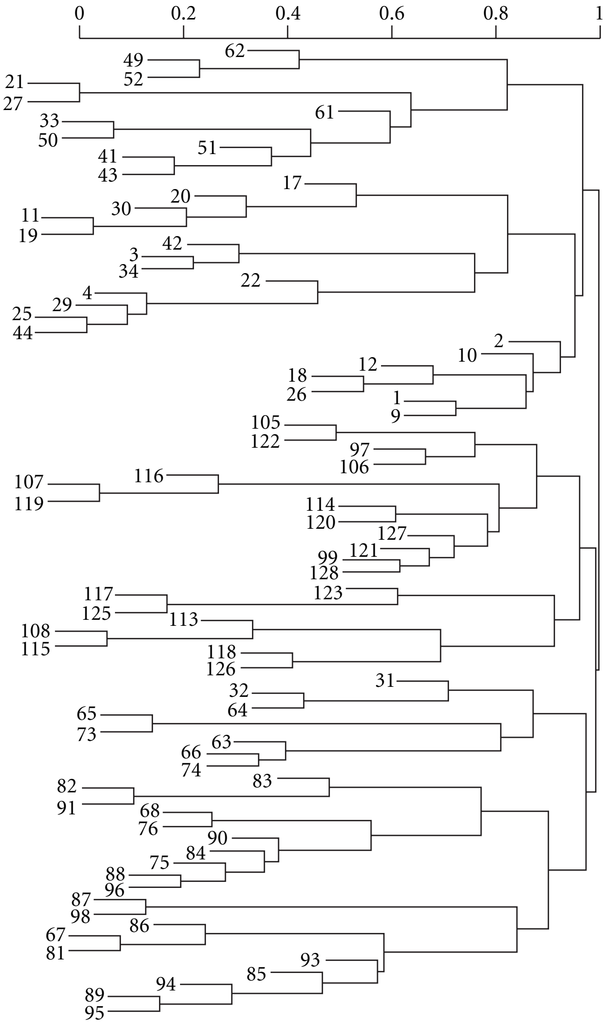


Figure 1: Dendrogram from clustering of lung cancer data.

dropped from our study. Under this restriction we only kept 80 combinations, leaving out a total of 1, 264 patients.

4.2. Setting of the Algorithm. To run our algorithm EACCD, we chose parameters as follows. The choice of N depends on the rate at which dis in (2) converges to p_{ij}. A large number should be chosen for N, and for this purpose we set $N = 10000$. Any theoretically possible choices of K was used in running PAM, and thus we set $K_1 = 2$ and $K_2 = 79$, due to availability of 80 objects. In addition, the log-rank test was

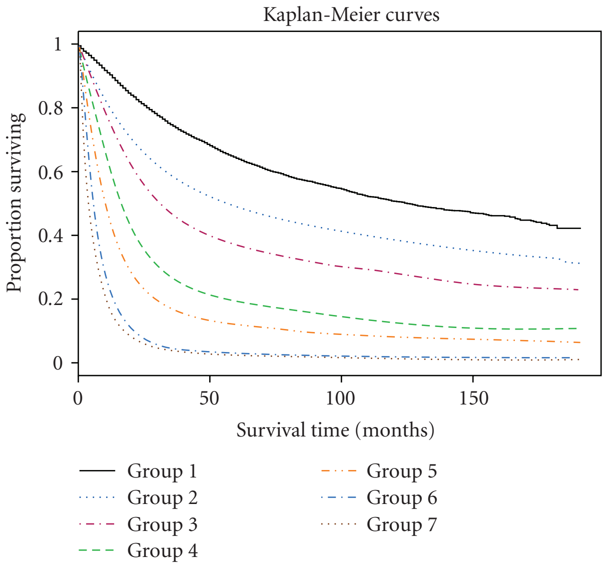


Figure 2: Survival curves of seven groups in Table 3.

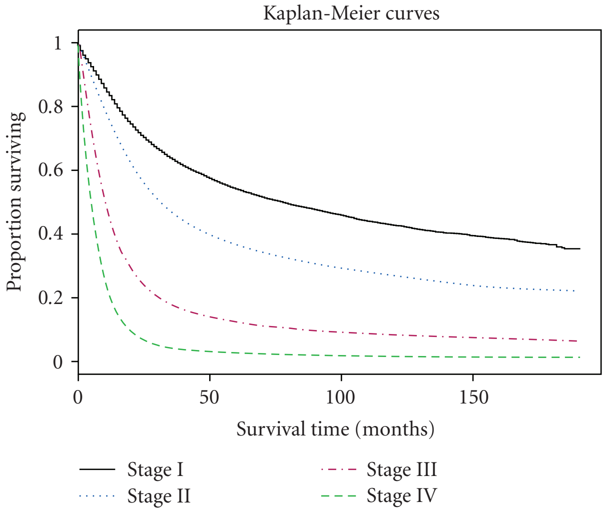


Figure 3: Survival curves of four TNM stages.

used to obtain the measure (1) for the PAM algorithm. And the average linkage was employed as a hierarchical clustering method.

4.3. Results from Cluster Analysis. The output of cluster analysis for these 80 combinations is shown in Figure 1, where for simplicity Comb has been removed from each combination or label. It is straightforward to use the dendrogram shown in Figure 1. Cutting off the dendrogram at a specified height of the dissimilarity axis partitions data into disjoint clusters or groups. Cutting at different heights usually leads to different numbers of groups. As an example, if we cut the dendrogram in Figure 1 at a height slightly above 0.90, then we obtain 7 groups shown in Table 3. The log-rank test shows that any two groups differ significantly (using a significance level of 0.01) in their survival functions. Figure 2 shows the Kaplan-Meier estimates of the survival curves for the 7 groups. These 7 groups and their survival curves constitute a prognostic system for lung cancer patients, as discussed in step (ii) of the Section of Introduction. Prediction using this system is then carried out in the usual way. In comparison, those 4 survival curves from the TNM system, based on all the patients from the 80 combinations, are provided in Figure 3.

Some observations come immediately from Table 3. Group 1, 5, 6, and 7 only contain some cases from Stage I, III, IV, and IV, respectively. Both groups 2 and 3 contain Stage I cancer cases, indicating that additional relevant parameters are associated with increased relative biologic aggressive tumor behavior. Group 4 consists of some cases from Stage I, II, and III, suggesting that localized biologically aggressive cancers may have the same survival as more indolent advanced staged cancers.

5. Conclusion

In this paper we have introduced an ensemble clustering based approach to establish prognostic systems that can be used to predict an outcome or a survival rate of cancer patients. An application of the approach to lung cancer patients has been given.

Generalizing or refining the work presented in this paper can be done in many ways. Our algorithm EACCD actually is a two-step clustering method. In the first step, a dissimilarity measure is learnt by using PAM, and in the second step, the learnt dissimilarity is used with a hierarchical clustering algorithm to obtain clusters of patients. These clusters of patients form a basis of a prognostic system. Improvement of dissimilarity measures (1) and (2), as well as the effect of different algorithms used in each step will be investigated in our future work. Refined algorithms, based on EACCD, will be sought and resulting prognostic systems with clinical applications will be reported. This constitutes our main research work in the future.

Acknowledgment

This work was partially supported by the National Science Foundation Grant CCF-0729080.

References

[1] SEER, http://seer.cancer.gov/.

[2] NCDB, http://www.facs.org/cancer/ncdb/index.html.

[3] F. L. Greene, C. C. Compton, A. G. Fritz, J. P. Shah, and D. P. Winchester, Eds., *AJCC Cancer Staging Atlas*, Springer, New York, NY, USA, 2006.

[4] D. Chen, K. Xing, D. Henson, and L. Sheng, “Group testing in the development of an expanded cancer staging system,” in *Proceedings of the 7th International Conference on Machine Learning and Applications (ICMLA ’08)*, pp. 589–594, San Diego, Calif, USA, December 2008.

[5] D. Chen, K. Xing, D. Henson, L. Sheng, A. M. Schwartz, and X. Cheng, “A clustering-based approach to predict outcome

in cancer patients," to appear in *International Journal of Data Mining and Bioinformatics*.

[6] K. Xing, D. Chen, D. Henson, and L. Sheng, "A clustering-based approach to predict outcome in cancer patients," in *Proceedings of the 6th International Conference on Machine Learning and Applications (ICMLA '07)*, pp. 541–546, Cincinnati, Ohio, USA, December 2007.

[7] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York, NY, USA, 1990.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, NY, USA, 2001.

[9] J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, New York, NY, USA, 2nd edition, 2003.

[10] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958.

[11] E. A. Gehan, "A generalized Wilcoxon test for comparing arbitrarily singly-censored samples," *Biometrika*, vol. 52, pp. 203–223, 1965.

[12] N. Breslow, "A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship," *Biometrika*, vol. 57, no. 3, pp. 579–594, 1970.

[13] R. E. Tarone and J. Ware, "On distribution free tests for equality of survival distributions," *Biometrika*, vol. 64, no. 1, pp. 156–160, 1977.

[14] A. L. N. Fred and A. K. Jain, "Data clustering using evidence accumulation," in *Proceedings of the 16th International Conference on Pattern Recognition (ICPR '02)*, vol. 4, pp. 276–280, Quebec, Canada, August 2002.

[15] D. Chen, Z. Zhang, Z. Liu, and X. Cheng, "An ensemble method of discovering sample classes using gene expression profiling," in *Data Mining in Biomedicine*, P. M. Pardalos, V. L. Boginski, and A. Vazacopoulos, Eds., vol. 7 of *Springer Optimization and Its Applications*, pp. 39–46, Springer, New York, NY, USA, 2007.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 357937, 5 pages
doi:10.1155/2009/357937

Research Article

Characterizing Gene Expressions Based on Their Temporal Observations

Jiuzhou Song,[1] Hong-Bin Fang,[2] and Kangmin Duan[3]

[1] *Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA*
[2] *Division of Biostatistics, University of Maryland Greenebaum Cancer Center (UMGCC), Baltimore, MD 21201, USA*
[3] *Department of Microbiology and Infectious Diseases, Health Sciences Centre, University of Calgary, Calgary, AB, Canada T2N 4N1*

Correspondence should be addressed to Jiuzhou Song, songj88@umd.edu

Received 31 October 2008; Revised 4 February 2009; Accepted 5 March 2009

Recommended by Zhenqiu Liu

Temporal gene expression data are of particular interest to researchers as they contain rich information in characterization of gene function and have been widely used in biomedical studies. However, extracting information and identifying efficient treatment effects without loss of temporal information are still in problem. In this paper, we propose a method of classifying temporal gene expression curves in which individual expression trajectory is modeled as longitudinal data with changeable variance and covariance structure. The method, mainly based on generalized mixed model, is illustrated by a dense temporal gene expression data in bacteria. We aimed at evaluating gene effects and treatments. The power and time points of measurements are also characterized via the longitudinal mixed model. The results indicated that the proposed methodology is promising for the analysis of temporal gene expression data, and that it could be generally applicable to other high-throughput temporal gene expression analyses.

1. Introduction

The high-throughput gene expression techniques, such as oligonucleotide and DNA microarray, serial analysis gene expression (SAGE) make it possible now to quickly generate huge amount of time series data on gene expression under various conditions [1–5], and have been widely applied in biomedical studies. The current temporal gene expressions usually have several main features: containing large scale of data set, having many genes, involving many procedure noises, and absenting statistical confidence, but few measuring time series levels. Using the difference at two or very few time points to understand changes has also some fundamental limitations. It tells us nothing about each gene's trajectory, and does not consider "overall" difference, nor does it allow studying evolution difference. For these such data with observations at very few time points, the current widely used analysis methods are various clustering methods, fold expression changes, ANOVA [6–9], and recently the hidden Markov chain models (Yuan and Kendziorski 2006). It is simple to interpret the results, and all the available data are analyzed when these methods are applied. However, there are problems associated with these methods which include merely qualifying characteristics of the gene behaviors and clearly absenting quantitative description, and it may take a risk of having false positive and false negative when looking strictly at fold change [9, 10]. Some genetic information may be lost using fold change analysis, and difficulties arise when genes having a bigger folds change in one expression experiment have different performance in multiple arrays or different experiments. It is even more problematic when multiple testing was carried out. For the widely used ANOVA or univariate method, it only analyzes difference between observed means and treats changes of individual gene profile as noise. The main limitation is that the data must be balanced, that is, all measurements occur at same times for all genes, no distinction between unequally spaced time points and equally spaced time points. The ANOVA does not produce a parameter that evaluates the rate of change over time for different treatment groups. Besides, it provides an oversimplification representation for the mean of a data set. The generalized linear models are also used in analyzing gene expression data, but they are based on analyzing the data at each time point separately. They do not take into account the fact that the gene expression measurements are not independent and do not address the difference in how the mean changes over time. Both the "classical" univariate and multivariate procedures assume that covariance matrix of each data is the same for all measurements at different times, regardless of group or compound symmetry. This assumption implies a very pattern of correlation among observations taken on the same unit at different times which is quite unrealistic for longitudinal data [11]. The other characteristic shared both by the classical univariate and multivariate methods is that time itself does not appear explicitly in the model.

By characterizing the entire pattern of gene expression, and distinguishing the individual gene profile changes subgroup and population-average profile changes, precise estimates with good capability and excellent combination of gene and condition effects were achieved with observations at much more time points. A prospective cohort study where repeated measures are taken over time for each gene is usually designed to answer the following two questions. First, how many observation points are needed over time? Second, how are the variables of interest including genes and conditions associated with each other over time? Therefore, the longitudinal observations with enough time points are most appropriate for the investigation of individual gene changes over time and for the study of effects of other factors such as experimental conditions. In this paper, we illustrate the strategy with an example of a 15-gene set in *Pseudomonas aeruginosa* expressed in three conditions and measured at 48 time points. These 15 genes are either quorum-sensing (QS) genes or quorum sensing regulated genes. Quorum sensing system is a bacterial gene regulatory system that employs small secreted molecules called autoinducers as signaling molecules to coordinate gene expression in a population manner. The autoinducers synthesized and diffused into the growth medium by individual cells increase in amount when the cell number increases, and when the concentration of autoinducers reaches a threshold they bind to cognate transcription regulator to modulate transcriptions of the bacterial genes. So the cell behaves as a whole. The quorum sensing systems in *P. aeruginosa* play a central role in regulating virulence factor expression and in biofilms formation. It has been reported that the expression of one of the genes in QS systems, rhlI is regulated by the iron conditions of the growth medium. However, the extent that this gene is regulated by iron availability is rather small. It is hard to assess the importance of this effect of iron on the QS system in *P. aeruginosa.* Employing the strategy described in this paper, we are able to determine the definite effect of iron availability using a relative large dataset which includes 15 genes over 48 time points in three different conditions totaling 2160 data points.

To analyze such data of temporal gene expressions, the longitudinal mixed model is used. The linear mixed models are extensions of linear regression models used to analyze longitudinal (correlated) data. They accommodate both fixed effects and random effects where the random effects are used to model between-gene variation and the correlation induced by this variation. Linear mixed models are extremely flexible analysis tools, which are especially suitable for unbalanced data with unequally spaced time points and of emphasis on both individual gene level and population-level components. The longitudinal mixed model analysis we present provides a strategy to analyze more complex time series gene expression datasets. The gene expression longitudinal data is characterized by repeated observations over time on the same set of genes, and the repeated observations on the same gene tend to be correlated, therefore, any appropriate statistical analysis must take this correlation into account. The longitudinal mixed model analysis is useful to identify general trends within genes over time, to detect nonlinear changes over time, and also to provide information about the amount of interindividual gene variability. This analysis incorporates different subgroups on the same graph to explain interindividual gene variability.

Table 1: Culture media.

Condition treatments	Description
C1T13	TSBDC
C2T13	TSBDC + 400 ug/mL EDDA
C3T13	TSBDC + 50 ug/mL $FeCl_3$

2. Materials and Methods

2.1. Gene Expression Data in P. Aeruginosa. The promoter regions of selected *P. aeruginosa* virulence factors were amplified by PCR using oligonucleotide primers synthesized [12] according to the PAO1 genome data and PAO1 chromosomal DNA as the template. The PCR amplified promoter regions were then cloned into the *Xho*I-*Bam*HI sites of pMS402 and transformed into PAO1 by electroporation. PCR and DNA manipulation and transformation were performed following general procedures. The promoterless *luxCDABE* operon in pMS402 enables the activity of the promoter fused upstream of the operon to be measured as counts per second (CPS) of light production in a Victor^2 multilabelcounter [12].

TSBDC minimal medium supplemented with EDTA (400 ug/mL) and 50 ug/mL $FeCl_3$ was used in gene expression assays (Table 1). Overnight cultures of the reporter strains were diluted 1 : 200 in a 96-well microtiter plate and the promoteractivity of the virulence factors in different conditions was measured every 30 minutes for 24 hours. Bacterial growth was monitored at the same time by measuring the optical density at 620 nm (OD_{620}) in the Victor^2 multilabel counter.

2.2. Statistical Methods. To analyze these longitudinal data of temporal gene expressions, the mixed model

$$Y_i = X_i\beta + Z_ib_i + \varepsilon_i \qquad (1)$$

will be used, where Y_i is an $(n_i \times 1)$ vector of expression for the ith gene, $i = 1, \dots, m$. X_i is an $(n_i \times p)$ design matrix that characterizes the systematic part of the gene expression, for example, depending on covariates and time. β is a $(p \times 1)$ vector of parameters usually referred to as fixed effects, that complete the characterization of the systematic part of the gene expression. Z_i is an $(n_i \times k)$ design matrix that characterizes random variation in the response attributable to among genes. b_i is a $(k \times 1)$ vector of the random effects variables that completes the characterization of the among-gene variation. ε_i is an $(n_i \times 1)$ vector of within-gene errors characterizing variation due to the way in which the expression levels are measured on the ith gene.

The data vector Y_i has a multivariate normal distribution with $E(Y_i) = X_i\beta$, $\text{var}(Y_i) = Z_iDZ_i' + R_i = \Sigma_i$, and $Y_i \sim N(X_i\beta, \Sigma_i)$. Here, the usual assumptions are $b_i \sim N(0, D)$, D is a $(k \times k)$ covariance matrix that characterizes variation due to among-gene source, and the dimension of D corresponds to the number of among-gene random effects in model. $\varepsilon_i \sim N(0, R_i)$, R_i is an $(n_i \times n_i)$ covariance matrix that characterizes variance and correlation due to within-gene sources. The form of Σ_i implied by the model has two distinct components, the first having to do with variation solely from among-gene sources and the second having to do with variation solely from within-gene sources. We used maximum likelihood (ML), restricted maximum likelihood (REML), and minimum variance quadratic unbiased estimation (MIVQUE0) to estimate the covariance parameters of the G and R, respectively.

In order to check the influence of temporal measurements for longitudinal mixed analysis, we further constructed a dataset of the same dimension and with the same covariates and factor values for which power is to be calculated. With F-test statistics, we calculated noncentrality parameter (ϕ) and degrees of freedom ν_1 and ν_2, then power is calculated as $P(F_{\nu_1,\nu_2,0} > F_C)$, F_C is a critical value. All analyses were implemented by SAS package.

3. Results and Analysis

3.1. The Trajectories of the Longitudinal Gene Expression Data. To validate the models for our data set, we plotted the expression profiles for all genes under different conditions. The trajectories of the 15-gene set are shown in Figure 1. From the figure, we can see that there is high degree of variations between genes. There are also correlation genes at different time points, and the correlation structure cannot be ignored in analysis. The expression trajectories of the genes change over time for all of the genes, and at a certain time point, the change rate for each gene is different from other time point and from that of other genes. From Figure 2, we can see that the trajectories of experimental treats are also changing over time, and the change rate varies from conditions.

3.2. Choice of and Assessing the Goodness-of-Fit Covariance Structure. In the longitudinal data, there are three sources of error in the residual, including serial correlation, measurement error, and random component. In order to use longitudinal mixed-model methodology, it is assumed that the data has a linear mean and a reasonable covariance structure. The reasonable covariance structure is a parsimonious covariance just enough to be estimated with available current data and yet rich to capture probable covariance between gene expression observations. The fitting information shown in Table 2 provides some statistics about the estimated mixed model. The log likelihood supplies the estimation information of covariance G and R in the mixed models. Akaike's information criteria (AIC) can be used to compare models with the same fixed effects but different variance structures. Models having the smallest AIC are deemed the best. The Schwarz Bayesian criteria (BIC) are also computed, and models with smaller BIC are also preferred. The six models with different covariance structure were fitted, and preference was selected based on the AIC and BIC values. Inspection of AIC and BIC values for each of the six models revealed that the values of both the AIC and BIC in the assumed same covariance structure are larger than those of the assumed different ones. Both criteria are the smallest for the chosen separate spatial power (SP) structures for each treatment. The values of both AIC and BIC in SP are the minimum among the models. The log likelihood of the model is also the best for separate SP structures. As both criteria agree, it would be sensible to choose the

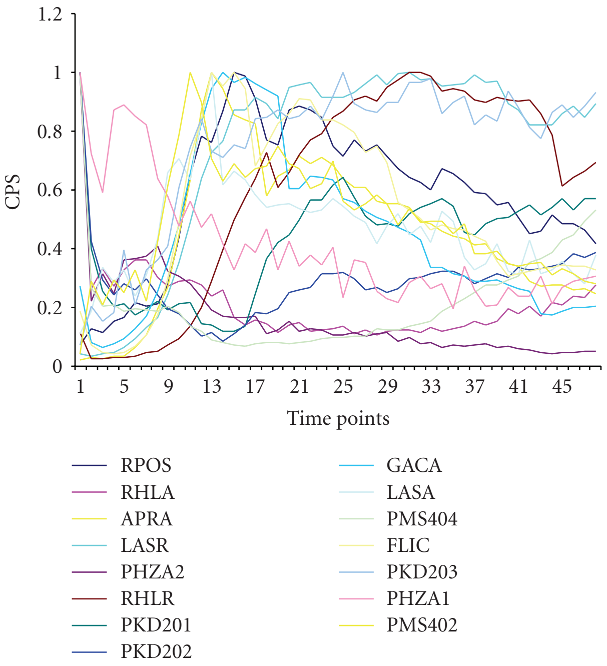


Figure 1: The trajectories of the 15 gene-set in C1T13: TSBDC condition.

Table 2: Covariance structures using ML.

Model	Description	AIC	BIC	−2 log likelihood
1	General linear model (GLM)	1811.8	1856.2	1798.8
2	Compound symmetry (CS)	1811.5	1856.0	1796.7
3	Variance components (VC)	1665.0	1651.3	1645.0
4	Heterogeneous CS (CSH)	1636.8	1618.0	1610.8
5	Spatial power (SP)	1689.2	1685.6	1600.2

AIC: Akaike's information criteria; BIC: Bayesian information criteria for each model selected.

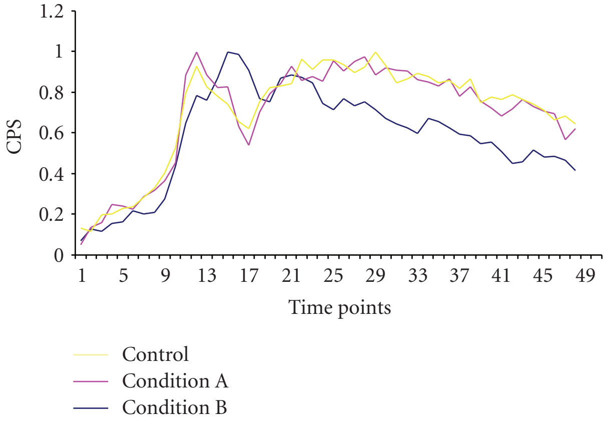


Figure 2: The trajectories of one gene in 3 conditions. Control: TSBDC, Condition A: TSBDC + 50 ug/mL $FeCl_3$ Condition B: TSBDC + 400 ug/mL EDDA.

model to represent the covariance structure that has different variance and covariance in different treatments. Interestingly, we found there were the almost same values AIC, BIC, and likelihood value between GLM and CS model, which indicated that univariate GLM calculations are identifical to MIXED estimates when using CS for the balanced data sets. The multivariate GLM cannot determine best fit when the data set is a longitudinal data.

3.3. Power and Sample Size Determination for Longitudinal Mixed Model. In statistical analysis, one typically expresses the belief that some effects exist in a population by specifying an alternative hypothesis to H_1, a null hypothesis H_0 as the assertion that effect does not exist and attempt to gather evidence to reject H_0 in favor of H_1. If H_0 is rejected but there is really no effect, this is called a Type I error, which is usually designated α; if there really is an effect in the population but H_0 is not rejected, then a Type II error has been made, which is usually designated β. The probability $1 - \beta$ of avoiding a Type II error, that is, correctly rejecting H_0 and achieving statistical significance, is called the power. We simulated our data structure and calculated the power of estimating condition effects via the longitudinal mixed model. As shown in Figure 3, we found the model can get maximum power while more than 7 or 8 measurements were taken. So the 48 temporal measurements of each gene in our research could have enough power to obtain the estimation of treatments and gene effects.

3.4. Estimation of the Effect of Iron Condition on QS Genes by the Mixed Model. We adopted the longitudinal mixed model with heterogeneous compound symmetry variance to estimate the effects of iron condition on QS genes expression. From Figure 2, the effects of the culture media TDBDC and TSBDC + 400 ug/mL EDDA are almost equal and higher than that of TDBDC + $FeCl_3$. Comparing with the TSBDC, the addition of TSBDC + 50 ug/mL $FeCl_3$ positively regulates the expression of these genes as shown in Figure 4. To check the detailed differences of the genes, the longitudinal mixed model was used to estimate the gene effects, as shown in Figure 5. We found that most of genes, including FliC, LasR, PKD202, PKD203, and PhlR, demonstrate positive expression effects in condition of addition of 400 ug/mL EDDA, whereas PhlA shows opposite expression effect.

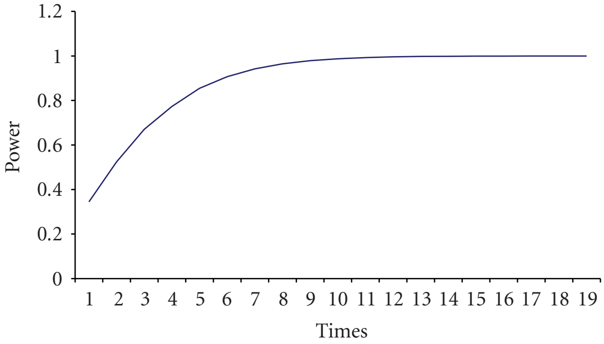


Figure 3: Power analysis under the longitudinal mixed model with heterogeneous compound symmetry variance structure.

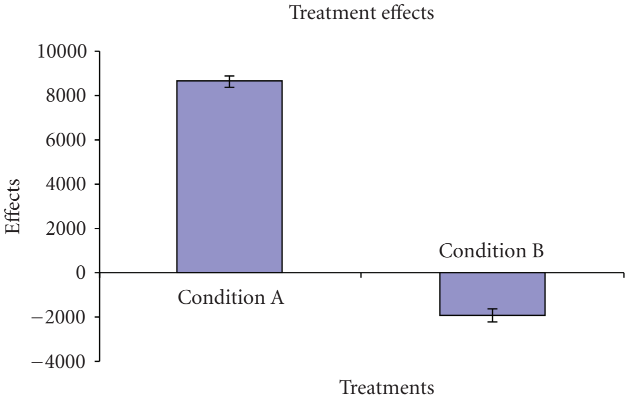


Figure 4: The estimation of condition effects. Condition A: TSBDC + 50 ug/mL $FeCl_3$,Condition B: TSBDC + 400 ug/mL EDDA.

4. Discussion

The identification of genes that show changes in expression between varying biological conditions is a frequent goal in microarray experiments. Under different biological conditions, the patterns of gene expressions may be various. To obtain efficient information for temporal gene expression, the number of longitudinal observations should be enough for individual gene changes over time and the study of effects with biological conditions.

In longitudinal studies, time effect is the changes over time for each gene, and cohort effect is the 22 differences among genes in their baseline values. Longitudinal studies can distinguish these time and cohort effects while cross-sectional studies cannot. In this paper, we have considered mixed model with longitudinal covariates, the analysis of longitudinal data should take into account firstly, the within-subject correlation, secondly the measurements taken at unequal time intervals and finally the missing observations. Repeated measures analysis of variance can be used to

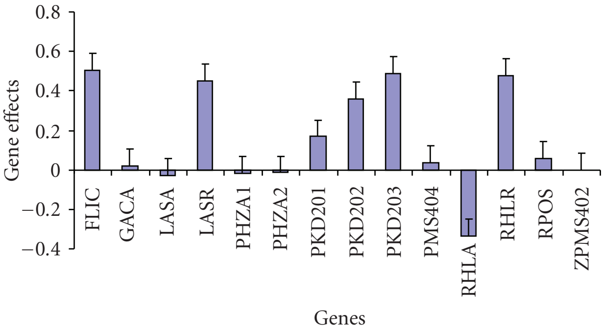


Figure 5: The estimation of gene effects under condition TSBDC + 50 ug/mL $FeCl_3$.

analyze longitudinal or repeated measures data for balanced study design, that is, when all genes are measured at equal time points and there are no missing data. In large scale of gene expression analysis, if having unbalanced datasets in longitudinal studies, it is necessary to use some alternative techniques which can handle unbalanced data. In this research, we confirmed that univariate GLM calculations are identifical to MIXED estimates when using CS for balanced data sets. The multivariate GLM cannot determine best fit when the data set is a longitudinal data. Therefore, the procedures of best fit mixed model include: (1) the choice of the model, (2) the choice of the variance-covariance structure (specifying the working correlation structure for each gene, e.g., independence, exchangeable, stationary, and autoregressive), (3) assessing the goodness-of-fit of the model, and (4) assessing the goodness-of-fit of the variance covariance structure.

Although the paper only analyzed the effects of three treatments and 15-gene effects, it proved that the longitudinal mixed model is a feasible method in dense temporal gene expression analysis. We found that the addition of TSBDC + 50 ug/mL $FeCl_3$ positively regulates the expression of these genes in our analysis. It has been reported that iron availability in the growth condition affects the expression of genes. However, the changes of expression are rather small. It is thus difficult to assess whether there is a pronounced effect of iron on the QS genes. Accordingly the current analysis method, using the mixed model described aforementioned a definite effect could be determined. A comprehensive understanding of biological processes requires the acquisition of expression data at different developmental stages, in different tissues and different treatment conditions with different organisms. The addition of time as a variable allows observation of the modulation of gene expression whether due to the regulation of development or the changing impact of a treatment condition. The expectation is that high-throughput gene expression analysis conducted in the higher dimensions of genes, conditions, tissues, and time as variables will help elucidate what the genes do, when, where, and how they are expressed as elements of an orchestrated system under the effects of perturbations and developmental processes, and we will explore the possibility of generalized mixed model in higher dimensions expression data [13–16].

References

[1] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

[2] H. Zhu, Y. Tang, L. Ivanciu, et al., "Temporal dynamics of gene expression in the lung in a baboon model of E. coli sepsis," *BMC Genomics*, vol. 8, article 58, pp. 1–23, 2007.

[3] R. J. Cho, M. J. Campbell, E. A. Winzeler, et al., "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, no. 1, pp. 65–73, 1998.

[4] P. T. Spellman, G. Sherlock, M. Q. Zhang, et al., "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[5] J. Bjarnason, C. M. Southward, and M. G. Surette, "Genomic profiling of iron-responsive genes in *Salmonella enterica* serovar Typhimurium by high-throughput screening of a random promoter library," *Journal of Bacteriology*, vol. 185, no. 16, pp. 4973–4982, 2003.

[6] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[7] K. Y. Yeung and W. L. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[8] H. Li, Y. Luan, F. Hong, and Y. Li, "Statistical methods for analysis of time course gene expression data," *Frontiers in Bioscience*, vol. 7, pp. a90–a98, 2002.

[9] S. Draghici, O. Kulaeva, B. Hoff, A. Petrov, S. Shams, and M. A. Tainsky, "Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays," *Bioinformatics*, vol. 19, no. 11, pp. 1348–1359, 2003.

[10] T. S. Tanaka, S. A. Jaradat, M. K. Lim, et al., "Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 16, pp. 9127–9132, 2000.

[11] S. L. Zeger and K.-Y. Liang, "An overview of methods for the analysis of longitudinal data," *Statistics in Medicine*, vol. 11, no. 14-15, pp. 1825–1839, 1992.

[12] K. Duan, C. Dammel, J. Stein, H. Rabin, and M. G. Surette, "Modulation of *Pseudomonas aeruginosa* gene expression by host microflora through interspecies communication," *Molecular Microbiology*, vol. 50, no. 5, pp. 1477–1491, 2003.

[13] J. Z. Song, K. M. Duan, T. Ware, and M. Surette, "The wavelet-based cluster analysis for temporal gene expression data," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, Article ID 39382, 7 pages, 2007.

[14] H. V. Westerhoff, E. Mosekilde, C. R. Noe, and A. M. Clemensen, "Integrating systems approaches into pharmaceutical sciences," *European Journal of Pharmaceutical Sciences*, vol. 35, no. 1-2, pp. 1–4, 2008.

[15] H. V. Westerhoff, A. Kolodkin, R. Conradie, et al., "Systems biology towards life in silico: mathematics of the control of living cells," *Journal of Mathematical Biology*, vol. 58, no. 1-2, pp. 7–34, 2009.

[16] I. P. Androulakis, E. Yang, and R. R. Almon, "Analysis of time-series gene expression data: methods, challenges, and opportunities," *Annual Review of Biomedical Engineering*, vol. 9, pp. 205–228, 2007.