University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2006 *Paper* 6

Nonparametric Pathway-Based Regression Models for Analysis of Genomic Data

Zhi Wei* Hongzhe Li†

*

†University of Pennsylvania, hli@cceb.upenn.edu

Nonparametric Pathway-Based Regression Models for Analysis of Genomic Data

Zhi Wei and Hongzhe Li

Abstract

High-throughout genomic data provide an opportunity for identifying pathways and genes that are related to various clinical phenotypes. Besides these genomic data, another valuable source of data is the biological knowledge about genes and pathways that might be related to the phenotypes of many complex diseases. Databases of such knowledge are often called the metadata. In microarray data analysis, such metadata are currently explored in post hoc ways by gene set enrichment analysis but have hardly been utilized in the modeling step. We propose to develop and evaluate a pathway-based gradient descent boosting procedure for nonparametric pathways-based regression(NPR) analysis to efficiently integrate genomic data and metadata. Such NPR models consider multiple pathways simultaneously and allow complex interactions among genes within the pathways and can be applied to identify pathways and genes within pathways that are related to variations of the phenotypes. These methods also provide an alternative to mediating the problem of a large number of potential interactions by limiting analysis to biologically plausible interactions between genes in related pathways. Our simulation studies indicate that the proposed boosting procedure can indeed identify relevent pathways and genes within pathways. Application to a gene expression data set on breast cancer distant matastasis identified that Wnt, apoptosis and cell cycle regulated pathways are more likely related to the risk of distant metastasis among lymph-node-negative breast cancer patients. We also observed that by incorporating the pathway information, we achieved better prediction for cancer recurrence.

Nonparametric Pathway-Based Regression Models for Analysis of Genomic Data

Zhi Wei[1] **and Hongzhe Li**[2,*]

[1,2] Genomics and Computational Biology Graduate Group , University of Pennsylvania School of Medicine, Philadelphia, PA 19104, U.S.A.

[2] Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6021, U.S.A.

[*] *email:* hli@cceb.upenn.edu

Address for correspondence:

Hongzhe Li, Ph.D.

Department of Biostatistics and Epidemiology

University of Pennsylvania School of Medicine

423 Guardian Drive - 920 Blockley Hall

Philadelphia, PA 19104-6021.

Telephone: (215) 573-5038, Fax: (215) 573-4865

E-mail: hli@ucceb.upenn.edu

SUMMARY. High-throughout genomic data provide an opportunity for identifying pathways and genes that are related to various clinical phenotypes. Besides these genomic data, another valuable source of data is the biological knowledge about genes and pathways that might be related to the phenotypes of many complex diseases. Databases of such knowledge are often called the metadata. In microarray data anlysis, such metadata are currently explored in post hoc ways by gene set enrichment analysis but have hardly been utilized in the modeling step. We propose to develop and evaluate a pathway-based gradient descent boosting procedure for nonparametric pathways-based regression (NPR) analysis to efficiently integrate genomic data and metadata. Such NPR models consider multiple pathways simultaneously and allow complex interactions among genes within the pathways and can be applied to identify pathways and genes within pathways that are related to variations of the phenotypes. These methods also provide an alternative to mediating the problem of a large number of potential interactions by limiting analysis to biologically plausible interactions between genes in related pathways. Our simulation studies indicate that the proposed boosting procedure can indeed identify relevant pathways and genes within pathways. Application to a gene expression data set on breast cancer distant metastasis identified that Wnt, apoptosis and cell cycle regulated pathways are more likely related to the risk of distant metastasis among lymph-node-negative breast cancer patients. We also observed that by incorporating the pathway information, we achieved better prediction for cancer recurrence.

KEY WORDS: Microarray; SNPs; Gradient descent boosting; Tree; Gene set enrichment analysis; Additive models.

1 Introduction

New high-throughput technologies are generating many types of high-dimensional genomic and proteomic data in biomedical research. Important examples include microarray gene expression data measuring mRNA transcripts of about 25,000 genes in cells and single nucleotide polymorphisms (SNPs) array data on genotypes of over 500K SNPs. One important application of such data is to identify genes, their interactions, and pathways that might be related to various clinically relevant phenotypes, such as risk of developing cancers or outcomes from cancer treatments. One great challenge in studying the relationship between genomic data and phenotypes is to deal with high-dimensionality of the data and to model complex interactions between genes. Many new statistical and computational methods have been or are still being developed to solve this problem of "curse of dimensionality." Important recent developments include support vector machine (SVM) (Vapnik, 1998) and random forest methods (Breiman, 2001), which have gained much popularity in building predictive models and in identifying genes that are related to clinical phenotypes.

One limitation of all these popular approaches is that the methods are developed purely from computational or algorithmic points without utilizing any prior biological knowledge or information. For many complex diseases, especially for cancers, much biological knowledge or pathway information is available from many years of intensive biomedical research. The large body of information is now available, primarily through databases on different aspects of the biological systems. Such databases are often called metadata, which means data about data. Examples of such metadata include the gene ontology (GO) database (Gene Ontology Consortium, 2001), the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000), and other pathways databases available on the internet (e.g., www.superarray.com, www.biocarta.com). Currently, information derived from metadata such as known biological knowledge has been used primarily to select promising candidates for genetic risk characteriza-

tion and for studying gene-gene and gene-environment interactions in genetic association studies. For microarray gene expression data, the most commonly used approach for pathway analysis is to identify pathways that are over-represented by differentially expressed genes. Some popular tools include GENMAPP, CHIPINFO, GOMINER and ONTO-TOOLS. Such gene set enrichment analyses (GSEA) are of course very informative and are potentially useful (Tian *et al.*, 2005) for identifying pathways that might be related to the disease phenotypes. However, such information has not been utilized in the modeling stage for identifying genes, their interactions, and pathways that are related to the phenotypes. In addition, such GSEA analysis considers pathways separately. Since many complex phenotypes are believed to be associated with activity levels of multiple pathways, new statistical methods are required to consider multiple pathways simultaneously and to allow complex gene-gene interactions within pathways.

We propose in this paper to develop and evaluate a novel gradient descent boosting procedure for nonparametric pathways-based regression (NPR) analysis in order to efficiently integrate genetic or genomic data and metadata. Our approach utilizes both statistical methods and biological knowledge in reducing the dimensionality of the problem and in building pathways-based regression models. Compared to GSEA analysis, our NPR model considers multiple potential pathways simultaneously. In such an NPR modeling framework, known biological pathways are treated as first level regression units, and the genes within the pathways are treated as the second level regression units, where the genomic data, such as the expression levels of genes or SNPs data in a given pathway, are used to characterize the activity of the pathways and the activity levels across many pathways are related to the phenotypes by a regression model. This provides a nice biological interpretation of the resulting regression models. In addition, the NPR also provides an alternative of mediating the problem of a large number of potential interactions by limiting analysis to biologically plausible interactions between genes in related pathways. In general, risk interactions are more plausible between genes involved in a physical interaction, found in the

same pathways, or involved in the same regulatory network (Carlson *et al.*, 2004).

Boosting was introduced in the machine learning literature by Freund (1995) and Freund and Schapire (1996) and has demonstrated great empirical success on a wide variety of specially high-dimensional prediction problems, including analysis of microarray gene expression data (Dettling and Buhlmann, 2003; Horton *et al.*, 2005; Li and Luan, 2005). From the perspective of numerical optimization on function space, Friedman (2001) proposed a gradient descent boosting (GDB) procedure and demonstrated that such a procedure can be regarded as a stage-wise fitting of the additive models. We propose an extension of Friedman's GDB procedure to perform GDB by pathways for fitting the proposed NPR models using regression trees as weak learners. We also provide scores for assessing the relative importance of the genes and pathways.

The rest of the paper is organized as follows: we first introduce the NPR models. We then present a general pathway-based GDB procedure for identifying such NPR models for both the logistic regression model and the Cox proportional hazards model. We present simulation studies and analysis of a breast cancer distance metastasis data set to demonstrate and evaluate the proposed methods. Finally, we give brief discussion of the methods and results.

2 Nonparametric Pathway-Based Regression Models

Suppose that we have K pathways whose activities may be related to the phenotype of interest. Assume that there are p_k genes involved in the kth pathway. We allow that some genes belong to multiple pathways and let p be the total number of genes involved in the K pathways and therefore $p \leq \sum_{k=1}^{K} p_k$. Suppose that we have n independent individuals and we let y_i denote the phenotype (can be continuous, categorical, or censored survival data) for the ith individual. For binary phenotype, let $y_i = 1$ if the ith individual has the phenotype and -1 otherwise. For censored survival outcome, let $y_i = (t_i, \delta_i)$, where t_i is time to event or censoring and δ_i is an event indicator. Let $x_{ij}^{(k)}$ be the genomic measurement of the jth gene in the kth pathway for

the ith patient, $x_i^{(k)} = \{x_{i1}^{(k)}, \cdots, x_{ip_k}^{(k)}\}$ be the vector of the genomic measures of the genes in the kth pathway for the ith patient, and let $x_i = (x_i^{(1)}, \cdots, x_i^{(K)})$ be the vector of the genomic measurements of all the p genes. Here the genomic measurements can be SNP data or gene expression data. Our goal is to relate the phenotype data Y to $X = \{X^{(1)}, \cdots, X^{(K)}\}$ in order to identify the pathways that are related to the phenotype and to identify genes and their interactions that determine the pathway activities.

Here we assume that the phenotype is related to the total activity level across multiple pathways through an additive model,

$$F(X) = \sum_{k=1}^{K} F_k(X^{(k)}), \tag{1}$$

where $F_k(X^{(k)})$ can be interpreted as the activity level associated with the kth pathway as determined by the genomic measurements of the p_k genes in this pathway. We assume that conditioning on the genes of the pathways, the pathway activities across the K pathways are additive. For example, for a binary phenotype such as disease status or normal versus cancerous tissues, we can assume a generalized linear model such as a logistic model for Y,

$$Pr(Y = 1) = \frac{\exp(2(F(X) + \gamma Z))}{1 + \exp(2(F(X) + \gamma Z))}, \tag{2}$$

where $Y = 1$ for diseased individual and $Y = -1$ for normal individual, Z is the vector of other patient-specific covariates which is modeled parametrically with coefficients γ. For the censored survival phenotype, we can assume that the hazard function at time t given the observed genomic data X is modeled as

$$\lambda(t|X, Z) = \lambda_0(t) \exp(F(X) + \gamma Z), \tag{3}$$

where $\lambda_0(t)$ is the baseline hazard function and Z is a covariate vector and γ is the corresponding risk ratio parameter.

The main motivation of these models is that we aim to model complex interactions between genes within pathways nonparametrically, rather than assume particular parametric forms for

functions $F_k(X^{(k)}$. We use the term "nonparametric pathway-based regression" to particularly emphasize this point, i.e., the genetic and pathways effects are modeled nonparametrically. It is obvious that without any constraints on the functions $F_k(X^{(k)})$, model (2) or (3) is not identifiable. In the next section, we propose a general pathway-based gradient descent boosting procedure to identify such NPR models with the particular form of (2) or (3).

3 A Pathway-Based GDB procedure for the NPR models

We propose to extend the GDB procedure of Friedman (2001) to obtain an additive model with the form of model (1) using regression trees (Breiman *et al.*, 1984) as base learners. Regression trees provide a flexible way of modeling dependency between responses and the predictors and have been widely used in the context of boosting methods (Friedman *et al.*, 2001; Friedman, 2001). We also propose several statistics for assessing the importance of the genes and pathways that are related to the phenotype of interest.

3.1 A pathway-based GDB procedure for the NPR models

The key idea of our proposed extension of the boosting procedure of Friedman (2001) is that instead of performing gradient boosting over all the p genes, we perform gradient descent boosting over genes in each of the K pathways separately. We first consider the case when no other covariates are included in model (2) or (3). Let $L(y_i, F(x_i))$ be a loss function for the ith observation, which depends on the type of the phenotype. For binary phenotype and model (2), the loss function can be defined as

$$L(y, F(x)) = \sum_{i=1}^{n} L(y_i, F(x_i)) = \sum_{i=1}^{n} \log(1 + \exp(-2y_i F(x_i))), y_i \in \{-1, 1\}. \quad (4)$$

This is also the loss function used by Friedman *et al.* (2001) for LogitBoost and by Friedman (2001) for his GDB procedure. For survival phenotype, the loss function can be defined as

negative of the partial likelihood based on model (3),

$$L(y,\delta,F(x)) = \sum_{i=1}^{n} L(y_i,\delta_i,F(x_i)) = -\sum_{i=1}^{n}\delta_i\{F(x_i) - \log(\sum_{j=1}^{n} 1_{\{y_j \geq y_i\}}\exp(F(x_j)))\}. \quad (5)$$

This loss function is used in Li and Luan (2005) and Gui and Li (2005).

Extending the GDB procedure of Friedman (2001), our proposed pathway-based GDB procedure for the NPR models involves the following steps:

A Pathways-based GDB Procedure for the NPR models

1. *Initialization*, $F^{(0)}(X) = 0, F_k^{(0)}(X^{(k)}) = 0, k = 1,\cdots,K.$

Repeat, for $m = 1$ to M (boosting steps) do:

2. *Calculating the gradients w.r.t. each function $F_k(X^{(k)})$ over observed samples,*

$$\tilde{y}_i = -\left[\frac{\partial L(y_i,F(x_i))}{\partial F_k(x_i^{(k)})}\right]_{F(X)=F^{(m-1)}(X)}, i = 1,\cdots,n, k = 1,\cdots,K.$$

3. *Fitting trees to the gradient vector using $x^{(k)}$, let $h_k(x_i^{(k)};a)$ be the base learner procedure,*

$$(a^{(k)},\beta^{(k)}) = argmin_{a,\beta}\sum_{i=1}^{N}[\tilde{y}_i - \beta h_k(x_i^{(k)};a)]^2, k = 1,\cdots,K,$$

Let $k^* = argmin_k \sum_{i=1}^{N}[\tilde{y}_i - \beta^{(k)}h_k(x_i^{(k)};a^{(k)})]^2$.

4. *Line search over ρ for the pathway k^* selected in step 3,*

$$\rho_m = argmin\sum_{i=1}^{n} L(y_i, F^{(m-1)}(x_i) + \rho h_{k^*}(x_i^{(k^*)};a^{(k^*)})).$$

5. *Updating the function with ν being the learning rate,*

$$F_{k^*}^{(m)}(X^{(k^*)}) = F_{k^*}^{(m-1)} + \nu\rho_m h_{k^*}(x_i^{(k^*)};a^{(k^*)}) \quad ,$$
$$F^{(m)}(X) = F^{(m-1)}(X) + F_{k^*}^{(m)}(X^{(k^*)}).$$

end For

end Algorithm

where M is the number of iterations, which serves as a shrinkage parameter and can be determined by cross-validation, $F^{(m)}(X)$ denotes the function $F(X)$ and $F_k^{(m)}(x^{(k)})$ denotes the function $F_k(x^{(k)})$ at the mth boosting step. Note that when $K = 1$, this algorithm reduces to the boosting

algorithm of Friedman (2001). In this algorithm, the gradients in step 2 of the generalized boosting algorithm are

$$\tilde{y}_{ik} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F_k(x^{(k)})}\right]_{F(X)=F_{m-1}(X)} = 2y_i/(1+\exp(2y_i F_{m-1}(x_i))) \tag{6}$$

for the logistic model (2)and

$$\tilde{y}_{ik} = \delta_i - \sum_{j=1}^{n} \delta_j 1_{\{t_i \geq t_j\}} \frac{\exp(F_m(x_i))}{\sum_{r=1}^{n} 1_{\{t_r \geq t_j\}} \exp(F_m(x_r))},$$

for the Cox model (3). Note that these gradients are the same for each different pathway k.

The key difference from the GDB algorithm of Friedman (2001) is found in step 3 and 4, where the calculation is done on a pathway by pathway basis. Step 3 aims to identify the pathway that gives the best fit of the negative gradients using the base learner. This effectively utilizes the known pathway information and reduces the dimensionality from considering all the genes to only considering those genes in a given pathway. In steps 4 and 5, the functions are updated by adding the tree corresponding to the k^*th pathway selected in Step 3. For many models, Step 4 simply reduces to a regression problem with a case weight.

In order to model interactions between genes in a given pathway, we propose to use a J-terminal node regression tree (Breiman *et al.*, 1984) as the base learning procedure in Step 3 of the algorithm for each pathway. For pathway k, each regression tree itself has the additive form

$$h_k(X^{(k)}; \{b_j^{(k)}, R_j^{(k)}\}_{j=1}^{J}) = \sum_{j=1}^{J} b_j^{(k)} I(X^{(k)} \in R_j^{(k)}),$$

where $\{R_j^{(k)}\}, j = 1, \cdots, J$ are disjoint regions that cover the space of all joint values of the variables $X^{(k)}$, and $a^{(k)} = \{b_j^{(k)}, R_j^{(k)}\}_{j=1}^{J}$ in the general boosting algorithm for the NPR models (Step 4). The boosting procedure with regression trees as base procedures inherits the favorable characteristics of trees such as robustness, and flexibility in modeling interactions (Breiman *et al.*, 1984). In addition, trees tend to be quite robust against the addition of irrelevant input variables and therefore serve as internal feature selection (Friedman, 2001; Breiman *et al.*, 1984). J controls the size of the tree, which is often chosen to be small.

Finally, if covariates Z are included in the NPR models (3) or (2), we can iterate between updating the parametric parameters γ by minimizing the loss function with $F(X)$ fixed and updating the nonparametric term $F(X)$ using the proposed boosting procedure.

3.2 Identification of important pathways and genes

We propose to apply cross-validation on the error rates for the logistic model (2) or the cross-validated partial likelihood to determine the number of boosting steps M. After M is determined and the function $F_k(X^{(k)})$ is estimated, we can address the issue of identifying relative importance of genes to the activity level of each pathway and identifying important pathways that are related to the phenotypes in the proposed NPR modeling framework. Although single trees are highly interpretable (Breiman *et al.*, 1984), the final function $F(X)$ identified by the pathways-based TGD procedure is a linear combination of trees and must therefore be interpreted in a different way. For a single tree T, Breiman *et al.* (1984) proposed to use

$$I_l^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 I(v(t) = l) \tag{7}$$

as a measurement of relevance for each predictor variable X_l for a tree with J nodes, where the sum is over the $J-1$ internal nodes, $I(.)$ is an indicator function and $v(t)$ is the splitting variable associated with the tth node. This score is basically the summation of the empirical improvement $\hat{i}_t^2$ in squared error risk as a result of a split at node t over the $J-1$ internal nodes of the tree. For models of additive tree expansions obtained from M boosting steps, Friedman (2001) suggested an importance score for the lth variable as

$$I_l^2 = \frac{1}{M} \sum_{m=1}^{M} I_l^2(T_m), \tag{8}$$

which is simply an average of the importance scores over the M trees obtained during the M boosting steps.

The importance scores defined by equations (7) and (8) can equally be applied to the additive trees obtained from the proposed pathways-based boosting procedure for the NPR models. However, for the final additive trees from the NPR model, we can in fact address more detailed questions about the role that genes and pathways play in determining the phenotypes. First, for each pathway k, we can assess the relevant influence of each gene j in this pathway by calculating the importance scores using the trees constructed based on the kth pathway, i.e.,

$$\hat{I}_{lk}^2 = \frac{1}{M_k} \sum_{m=1}^{M_k} I_l^2(T_{mk}),$$

where M_k is the number of times that the k pathway was selected in Step 3 of the proposed pathway-based boosting algorithm, and T_{mk} is the mth tree built based on the kth pathway. Second, the average of importance scores for genes selected within a pathway, which we call the pathway importance score, can be used as a measure of importance of this pathway to the phenotype. As in Friedman (2001), the most influential variable or pathway is given a score of 1, and the estimated importance scores of others are scaled accordingly.

4 Simulation Studies

In order to evaluate the performance of NPR's ability to identify important pathways and genes, we designed the following simulation studies, mimicking different possible biological scenarios. We assume that there are 50 candidate SNPs, denoted by $X_1, \ldots, X_{50}$, where SNPs $X_{(k-1)*10+1}, \ldots, X_{(k-1)*10+10}$ belong to the kth pathway, for $k = 1, \cdots, 5$. We generate X_is independently from Bernoulli distribution with probability of 0.25 of being 1. We generate disease status variable Y based on the following logistic regression model,

$$Pr(Y = 1|X) = \frac{\exp(2F(X))}{1 + \exp(2F(X))} \tag{9}$$

where $Y = 1$ for disease and $Y = -1$ for disease-free. We consider four different models with the following four predictive functions,

$$\begin{aligned}
F_1(X_1, \ldots, X_{50}) &= -0.6 + 0.25X_1 + 0.25 * X_2 + 0.5(X_1 * X_2), \\
F_2(X_1, \ldots, X_{50}) &= -0.5 + 0.5(X_1 * X_2), \\
F_3(X_1, \ldots, X_{50}) &= -0.5 + 0.5((X_1 * X_2) \text{ OR } (X_{11} * X_{12})), \\
F_4(X_1, \ldots, X_{50}) &= -0.5 + 0.5((X_1 * X_2) \text{ OR } (X_1 * X_{12})),
\end{aligned}$$

where in function F_3 and F_4, the OR operator returns value 1 if at least one of the two product terms is 1. Among these four models, model 1 presents the standard logistic regression model with two SNPs involved, model 2 assumes that only when there are two mutations on SNP1 and SNP2 from pathway 1 does the disease risk increase, model 3 assumes that there are two independent pathways involved; and model 4 also assumes that there are two pathways involved in disease risk; however, it assumes that SNP1 is involved in both pathways. For each model, the estimated disease rate is about 30%. We simulate data sets of 500 individuals and for each model, we repeat the simulation 100 times.

In the following analysis, we use the tree of depth three (i.e., at most three terminal nodes) as the base leaner procedure, which allows for two-way interactions between the variables. Since models 2-4 include only interaction effects, one would expect that the variable that entered in the tree at the later stage has a higher improvement score than those entered before. In this case, we adjust the important scores so that the two variables have the same importance scores.

4.1 Identification of the pathways

The four plots of Figure 1 show the frequencies during the pathway-based boosting procedure in which each of the five pathways was selected. It is clear that for models 1 and 2, pathway 1 was selected very frequently, and for models 3 and 4, both pathways 1 and 2 were selected almost

equally, indicating the importance of these pathways to the risk of the disease. Similarly, the four plots of Figure 2 show the boxplots of the relative importance scores of the five pathways over 100 replications. We observed that the relative importance scores are higher for pathway 1 for models 1 and 2, and are higher for pathways 1 and 2 for models 3 and 4, indicating that the pathway relative scores can indeed reveal which pathways are relevant to the disease risk.

4.2 Identification of the genes

To evaluate how well the proposed importance scores can be used for identifying genes that are related to the risk of disease, Table 1 shows the percentage of the true SNPs appearing in the top scoring variables over the 100 replications. For example, SNP1 and SNP2 are the SNPs with the first or second highest scores in 81% and 82% of the simulations for model 1 and 74% and 75% of the simulations for model 2. Similarly, for model 3, the relative importance scores for SNP1, SNP2, SNP11 and SNP12 are in general higher than the other SNPs. Among the 100 replications, the SNP1 and SNP2 are among the top four SNPs with the highest scores in 71% and 71% of the simulations, and SNP11 and SNP12 appeared among the top 4 SNPs in 59% and 58% of the simulations respectively. The SNP1 appeared among the top 3 SNPs in 84% of the replications for model 4. In addition, for model 4, SNP2 and SNP12 appeared among the top 3 SNPs with the highest scores in 69% and 60% of the replications, respectively. These numbers indicate that the relative importance scores can indeed capture the importance of the variables in the estimate of the function $F(X)$.

Figure 3 shows the boxplots of the importance scores for each of the 50 variables over 100 replications, indicating that the scores for the true SNPs are much higher than the other variables in most of the replications. We can clearly see that for model 1 and model 2, the relative importance scores for SNP1 and SNP2 are in general much higher than the other SNPs. Similarly for model 3, the importance scores for SNP1, SNP2, SNP11 and SNP12 are higher. For model

4, the importance score for SNP 1 is almost always the highest over 100 replications, indicating the importance of this SNP.

4.3 Comparison to other methods

As a comparison, we also performed analyses on the simulated data sets using the gradient descent boosting procedure of Friedman and the popular support vector machine method for feature selection as implemented in the program package GIST (*http://microarray.cpmc.columbia.edu/gist*). Neither of these two methods tried to utilize the pathway information. Table 1 shows the percentage over 100 simulations that the relevant SNPs were identified by these two methods. It is clear that the NPR methods tend to select the relevant SNPs more frequently than these two methods and the improvement is substantial for models 2, 3 and 4. For model 1, which is the standard logistic regression model including both main effects and interaction, the SVM seemed to select the SNP1 slightly better than the NPR method, but the difference is not significant. In addition, we also observed that the relative importance scores for the relevant variables obtained from the Friedman's procedure and the SVM are not as large as those obtained from the NPR. This comparison demonstrated the advantage of the pathway-based boosting procedure for the NPR models in selecting relevant variables, especially when the models do not follow the standard logistic regression models.

5 Application to lymph-node negative primary breast cancer data set

Wang *et al.* (2005) reported large Affymetrix-based gene expression profiling for 286 patients with lymph-node-negative primary breast cancer. These patients were treated between 1980-1995 with age at surgery ranging 26-86 and a median age at surgery of 52 yrs. No patient received

any adjuvant therapy. During the follow-up period, 180 of these patients were relapse-free at 5 yrs, and 106 of them developed distant metastasis. Gene expression profiling using Affymetrix HG-133A was performed on all these patients, including 17,819 transcripts that were present in two or more samples. We merged the Affymetrix probe IDs with SuperArray cancer related pathways/genes (www.superarray.com) and identified a subset of 245 genes in 33 cancer-related sub-pathways (see Table 2 for the pathways and the number of genes in each pathway). In addition, a set of 188 cancer-related genes is also included in our analysis. The numbers of genes within the pathways range from 2 (e.g., cell-cell adhesion and notch signaling pathways) and 81 genes (e.g., regulation of cell cycle).

We first performed the analysis using the logistic regression model (2). Using 10-fold cross validation on misclassification error rates, we chose the number of boosting steps to be 75, which gives an optimal misclassification error rate of 0.29. The left plot of Figure 4 shows the pathways with high relative scores and also high frequencies that were selected during the boosting procedure. We found that the Wnt pathway, the pathways related to apoptosis and cell cycle, and regulation of cell cycle are most likely related to the risk of distant metastasis.

Under the same 10-fold cross validation partitions, we performed analyses using several other well-known classifiers, including Random Forest, Bagging, Neural Network, BayesNet, Naive Bayes, Decision Stump, Ada Boosting M1, Logistic regression using the Weka software package (*http://www.cs.waikato.ac.nz/ml/weka/*) and SVM using the program GIST. The misclassifications that result from various procedures are shown in Table 3. It is clear that the NPR outperforms almost all of the competitors. This indicates that the pathways and genes selected by the NPR procedure may indeed be related to the risk of distant metastasis in lymph-node-negative breast cancer patients.

As a comparison, we also performed the analysis using the Cox model (3) with time to cancer relapse as the outcome. The right plot of Figure 4 shows the pathways with high importance

scores and high frequencies of being selected during the boosting procedure. The pathways identified are quite comparable to those identified using the logistic regression model.

6 Discussion

As the large body of biological information on various aspects of the biological systems and pathways is available through databases or metadata, it is important to utilize the information in modeling genomics data, especially in identifying genes and their interactions and pathways that might be related to the phenotypes. In this paper, we have introduced a nonparametric pathway-based regression model and proposed to extend the gradient boosting procedure of Friedman (2001) to obtain fits of such models. In addition, we have defined relative importance scores for genes within pathways and relative pathway importance scores in order to identify genes and pathways that might be related to the phenotypes. We have demonstrated the applications of such NPR models using both simulations and analysis of a breast cancer data set. Different from the traditional regression analysis, the proposed methods naturally incorporate biological pathways information. Different, also, from the commonly used gene set enrichment analysis, our method considers multiple pathways simultaneously and can easily incorporate other covariates.

The ensemble methods have been proposed mainly for predictive purposes, however, as demonstrated by Breiman (2001) and Friedman (2001) and also by our simulations, these methods can also be used for identifying variables that are relevant to the phenotypes. Although the interpretation of the resulting model is not as easy as that obtained from the traditional logistic regression or Cox regression, such models are more flexible and require fewer assumptions of the genetic effects. Although the relative importance scores used in this paper seem to perform well for identifying relevant variables, much future research needs to be done to rigorously investigate the problem of defining variable importance in the setting of ensemble methods. For example, important future research should assess the statistical significance of such importance scores,

by using bootstrap or permutations. In addition, it is also important to develop methods for identifying interactions among the variables based on the resulting ensemble of trees. Jiang and Owen (2002) proposed to apply a quasi-regression idea (An and Owen, 2001) for identifying the components based on the black-box functions. Similar quasi regression might be developed for the NPR models to identify important genes and pathways.

Another important issue that deserves further investigation is the sensitivity of the proposed methods to the misspecification of the pathways information and misspecification of the model. The first type of misspecification is that the genes included in the pathways do not really belong to the pathways. However, this should not create a big problem since these wrongly included genes should not be selected by the proposed methods. Another type of miss-specification is that the related genes are not included in the respected pathways. The third type of misspecification is that the relevant pathways are not included in the model. However, it should be noted that all types of regression analysis have such potential misspecification of the models. In defining our NPR model, we assume that genes within a pathway can interact; however, the pathways activities affect the phenotype in an additive model conditioning on the genes that the pathways include.

In summary, we have proposed a regression framework for identifying pathways and genes that are related to clinical phenotypes. The model and the pathway-based boosting procedure can be extended to include pathway specific gene-environment interactions to allow the same environmental risk factors to interact differently with different pathways. As more genes and pathways are being identified, we expect to see more applications of the proposed methods and its future extensions in analysis of genomic data.

Acknowledgments

This research was supported by NIH grant ES009911 and a grant from the Pennsylvania Department of Health. We thank Mr. Edmund Weisberg, MS at Penn CCEB for editorial assistance.

References

An J, Owen AB (2001): Quasi Regression. *Journal of Complexity*, 17:588-607.

Breiman L (2001): Random forests. *Machine Learning*, 45:5-32.

Breiman L, Firedman JH, Olshen RA, Stone C (1984): *Classification and Regression Trees*. Wadsworth.

Carlson CS, Eberle MA, Kruglyak L and Nickerson DA (2004): Mapping complex disease loci in whole-genome association studies. *Nature*, 429:446-452.

Dettling M and Buhlmann P (2003): Boosting for tumor classification with gene expression data. *Bioinformatics* 19, 1061-1069.

Freund Y (1995): Boosting a weak learning algorithm by majority. *Information and Computation*, 121: 256-285.

Freud Y and Schapire R (1996): Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, 148-156.

Friedman (2001): Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29: 1189-1232.

Friedman J, Hastie T and Tibshirani R (2000): Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28: 337-407.

The Gene Ontology Consortium (2000): Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25: 25-29.

Horton T, Dettling M and Buhlmann P (2005): Ensemble methods of computational inference. pp#293. In Gentleman R, Carey VJ, Huber W, Irizarry RA and Dudoit S eds *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer.

Jiang T and Owen AB (2002): Quasi-regression for visualization and interpretation of black box functions. Technical report, Department of Statistics, Stanford University.

Kanehisa M and Goto S (2002): KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28: 27-30.

Li H and Luan Y (2005): Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics*, 21: 2403-2409.

Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane I and Park P (2005): Discovering statistically significant pathways in expression profiling studies. *Proceedings of National Academy of Sciences*, 103: 13544-13549.

Vapnik V (1998): *Statistical Learning Theory*. Wiley.

Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA (2005): Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365: 671-9.

Table 1: Simulation results: percentage of the true variables that are among the genes with highest top two scores (for model 1 and 2), top four scores (for model 3) and top three scores (for model 4) over 100 replications. NPR: proposed pathway-based boosting procedure for the NPR models; GDB: the gradient boosting procedure of Friedman; SVM: support vector machine.

	Model 1		Model 2		Model 3				Model 4		
Method	X1	X2	X1	X2	X1	X2	X11	X12	X1	X2	X12
NPR	.81	.82	.74	.75	.71	.71	.59	.58	.84	.69	.60
GDB	.66	.68	.41	.47	.52	.59	.42	.44	.76	.44	.39
SVM	.90	.82	.32	.42	.45	.45	.41	.38	.81	.26	.26

Table 2: Pathways considered in breast cancer data analysis, including the numbers of genes in each pathway and a description of the pathways. The last set includes 188 genes that do not belong to a particular pathway.

Pathway ID	# of Genes	Description
1	18	Anti-apoptosis
2	4	VHLCaspase activation
3	3	DNA damage response
4	24	Factors involved in other aspect of apoptosis
5	8	Induction of apoptosis
6	10	Induction of apoptosis by signals
7	6	Regulation of apoptosis
8	3	Apoptosis others
9	13	Cell cycle arrest
10	4	Cell cycle checkpoint
11	29	Factors involved in other aspect of cell cycle
12	81	Regulation of cell cycle
13	6	Cell differentiation cell fate determination
14	63	Cell growth and/or maintenance
15	41	Cell proliferation
16	11	Growth factors
17	46	Regulation of cell proliferation differentiation growth and volume
18	10	Cell migration and motility
19	2	Cell-cell adhesion
20	6	Cell-matrix adhesion
21	10	Metalloendopeptidases (MMPs) and MMP inhibitors
22	13	Cell surface receptor linked signal transduction
23	9	Frizzled and frizzled-2 Signaling Pathways
24	17	G-protein coupled receptor protein signaling pathway
25	2	Insulin receptor signaling pathway
26	4	integrin-mediated signaling pathway
27	29	Intracellular signaling cascade
28	6	JAK-STAT cascade
29	2	Notch signaling pathway
30	3	RAS protein signal transduction
31	4	Rho protein signal transduction
32	13	Small GTPase mediated signal transduction
33	16	Wnt receptor signaling pathway
34	188	Other cancer-related genes

Table 3: Comparision of the average misclassification error rates of the NPR method and nine commonly used procedures based on 10-fold cross-validation for the breast cancer data set.

Classifiers	10-fold error rate
Random Forest	.33
Decision Stump	.42
Logistic Regression	.36
Neural Network	.29
Naive Bayes	.34
Bayes Net	.40
Bagging	.35
Ada Boost(M1)	.33
SVM	.42
NPR	.29

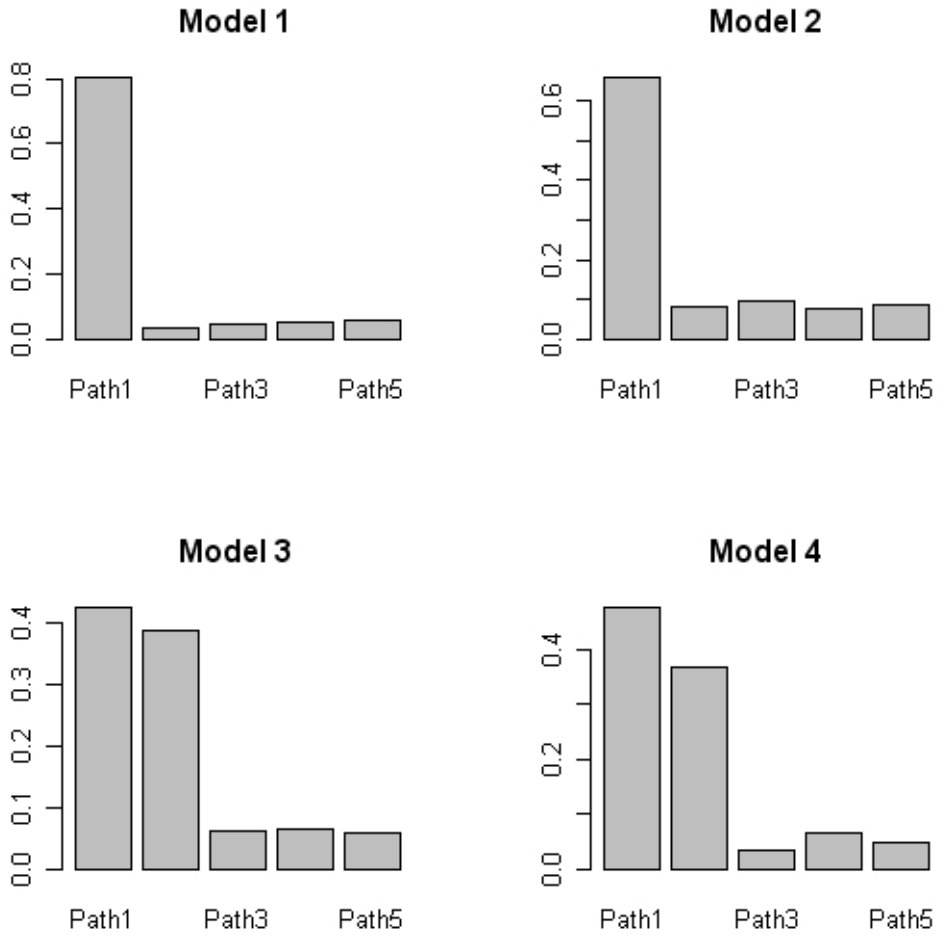

Figure 1: Simulation results: frequencies of the pathways selected during the pathway-based boosting procedure over 100 replications for each of the four simulated models.

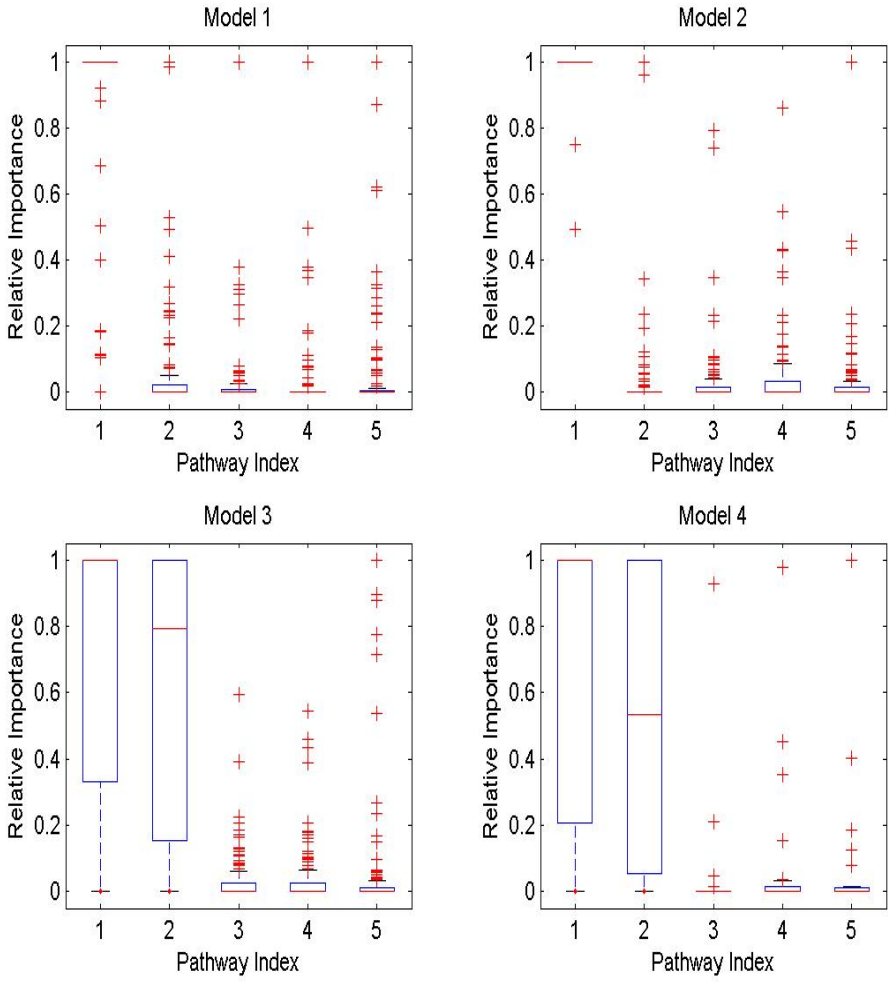

Figure 2: Simulation results: the boxplots of the relative importance scores for the four pathways based on 100 replications for each of the four simulated models.

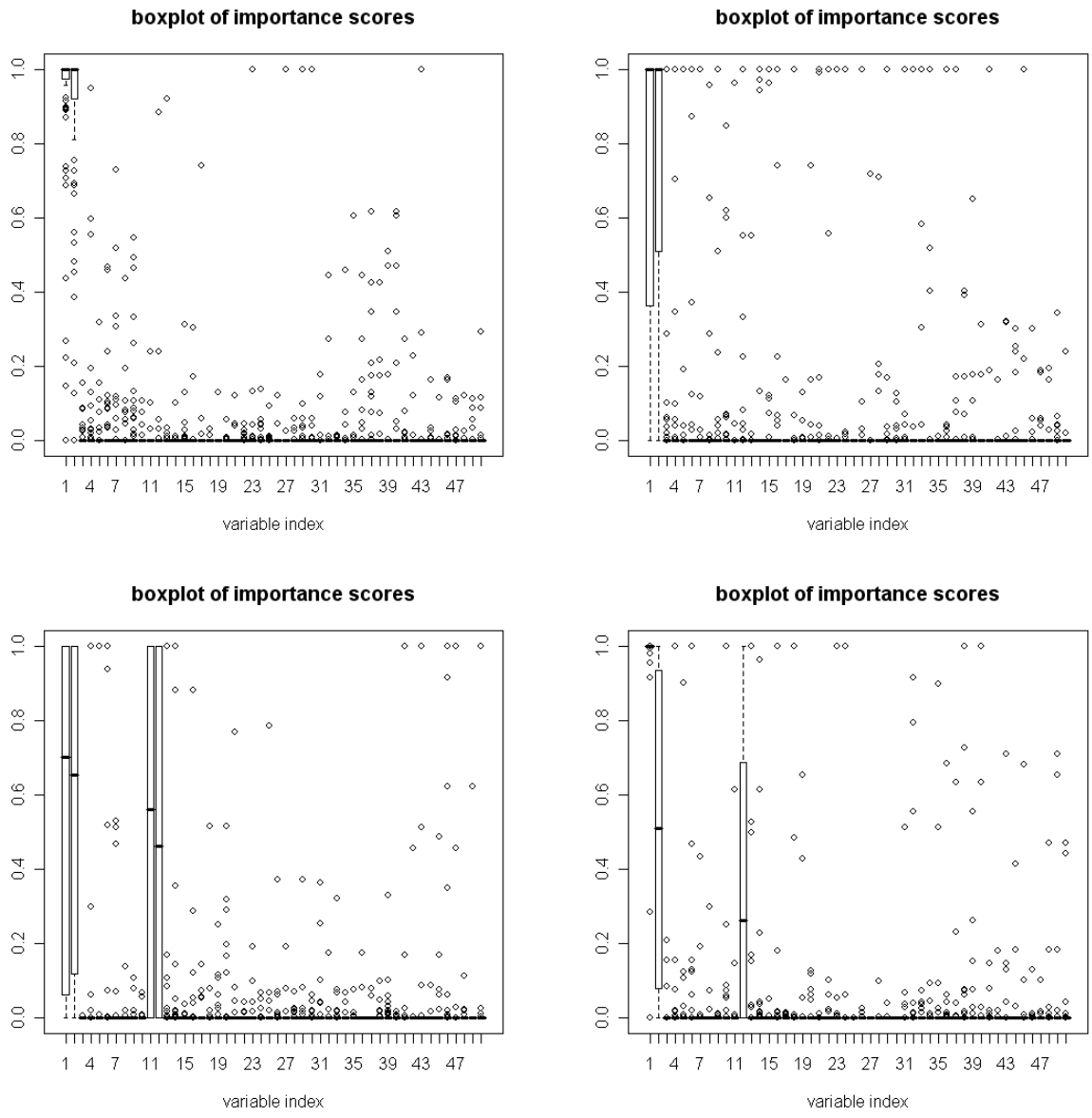

Figure 3: Simulation results: the boxplots of the relative imporance scores from the NPR models for each of the 50 variables over 100 replications.

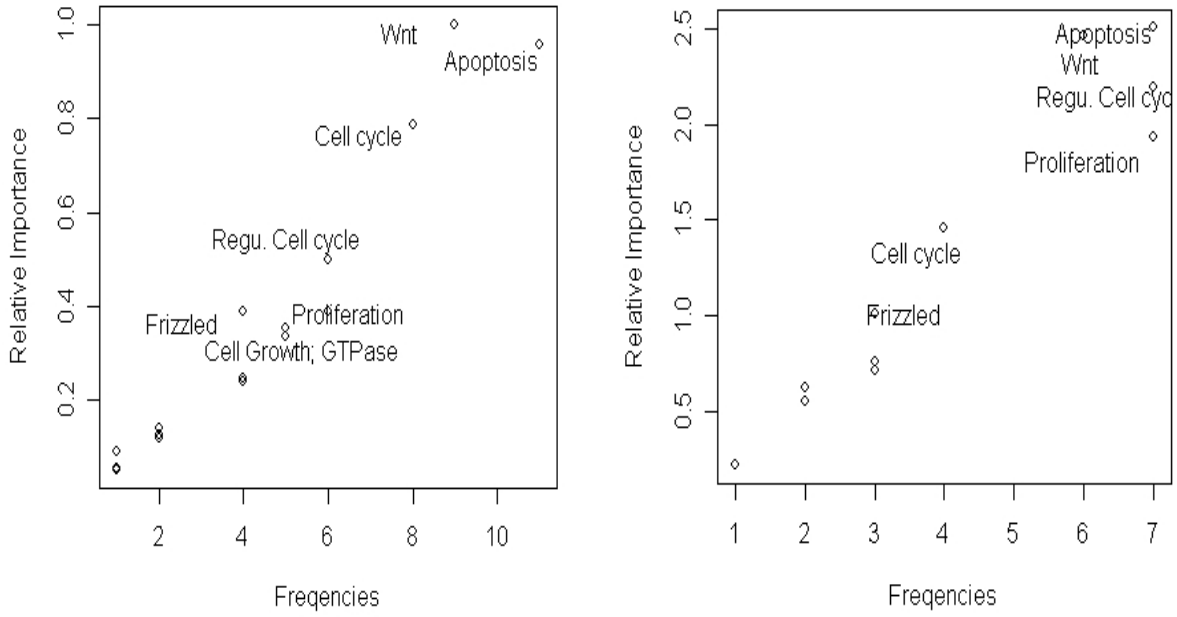

Figure 4: Results from analysis of breast cancer data set: plot of the frequencies of the pathways selected and the pathway importance scores during the boosting procedure for the logistic regression model (left plot) and the Cox model (right plot).