Nonparametric pathway-based regression models for analysis of genomic data

ZHI WEI
Genomics and Computational Biology Graduate Group, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

HONGZHE LI∗
Genomics and Computational Biology Graduate Group, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA and Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6021, USA
hli@cccb.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 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 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. 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. Results from analysis of other two breast cancer gene expression data sets indicate that the pathways of Metalloendopeptidases (MMPs) and MMP inhibitors, as well as cell proliferation, cell growth, and maintenance are important to breast cancer relapse and survival. We also observed that by incorporating the pathway information, we achieved better prediction for cancer recurrence.

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

∗To whom correspondence should be addressed.

© The Author 2006. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.
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 in machine learning literature, including support vector machines (SVM) (Vapnik, 1998) and random forest methods (Breiman, 2001), have gained much popularity in building predictive models and identifying genes that are related to clinical phenotypes (Furey and others, 2000; Guyon and others, 2002; Dettling and Buhlmann, 2003; Horton and others, 2005; Li and Luan, 2005).

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 database (The Gene Ontology Consortium, 2001), the Kyoto encyclopedia of genes and genomes database (Kanehisa and Goto, 2002), 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 characterization and for studying gene–gene and gene–environment interactions in genetic association studies. One example of such a pathway is the protein-directed growth and developmental Wnt signaling pathway (Hulsken and Behrens, 2000; Brown, 2001). The Wnt proteins form a large family of cell-secreted factors that control diverse aspects of development in mammals. The principal intracellular signaling pathway activated by these proteins has been elucidated. A simplified graph of the canonical Wnt/β-catenin signaling pathway given in Brown (2001) can be summarized as in Figure 1. In this pathway, Wnt proteins bind to receptors composed of a frizzled protein and either of the low-density lipoprotein receptor-related proteins LRP5 or LRP6. Signaling via Disheveled and/or Axin then results in inactivation of a multiprotein complex including Axin, adenomatous polyposis coli (APC), and glycogen synthase kinase-3β that normally renders β-catenin unstable. By inhibiting this complex, Wnt signals lead to accumulation of β-catenin in the cytosol and its entry into the nucleus. Once in the nucleus, β-catenin binds to proteins of the T-cell factor/lymphoid enhancer factor-1 family and modulates the expression of several target genes (Brown, 2001). It is now known that this pathway includes proto-oncogene products, such as β-catenin (marked as ∗ in the graph) and tumor suppressor proteins such as APC (marked as ∗∗ in the graph). Several polymorphisms in genes in this pathway and the abnormal expression of certain enzymes are shown to increase the breast cancer risk or risk of recurrence (Brown, 2001), indicating the importance of studying cancer from the aspect of cancer-related pathways.

For microarray gene expression data, the most commonly used approach for pathway analysis is to identify pathways that are overrepresented 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 potentially useful (Tian and others, 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 (GDB) 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 model allows potential interactions between genes within pathways, but not across pathways, which may provide an alternative of mediating the problem of a large number of potential interactions. In general, gene–gene interaction effects on phenotypes are more plausible between genes involved in a physical interaction, found in the same pathways, or involved in the same regulatory network (Carlson and others, 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 especially high-dimensional prediction problems, including analysis of microarray gene expression data (Dettling and Bühlmann, 2003; Horton and others, 2005; Li and Luan, 2005). Much has been written about the success of boosting procedures in producing accurate classifiers. The original boosting of Freund and Schapire (1996) works by sequentially applying a classification algorithm to reweighted versions of the training data, and then taking a weighted majority vote of the sequences of classifiers produced. Many papers have explored the use of simple tree-based classifiers (Breiman and others, 1984) as base classifier and have demonstrated that boosting consistently produces significantly lower test error rates than single decision trees. Friedman and others (2000) provided an elegant statistical justification of the boosting procedure and showed that boosting can be viewed as an approximation of additive modeling on the logistic scale using maximum Bernoulli likelihood as a criterion. This important insight opened a new perspective for using boosting in contexts other than classification. From the perspective of numerical optimization on function space, Friedman (2001) proposed a GDB procedure and demonstrated that such a procedure can be regarded as a stage-wise fitting of the additive models. Depending on the choice of the weak learner or base procedure, many different types of additive functions can be constructed. Within the boosting literature, the most commonly used base learner is a regression tree (Breiman and others, 1984), which renders final model as a linear combination of a large number of trees. Regression trees provide a natural nonparametric framework for modeling higher-order interactions between variables and the interaction order of the function to be estimated can be controlled by limiting the sizes of individual trees during the boosting procedures. 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 three breast cancer microarray gene expression data sets to demonstrate and evaluate the proposed methods. Finally, we give brief discussion of the methods and results.

2. NPR 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 \( k \)th 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 \( i \)th individual. For binary phenotype, let \( y_i = 1 \) if the \( i \)th 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 \( \delta_i \) is an event indicator. Let \( x_{ij}^{(k)} \) be the genomic measurement of the \( j \)th gene in the \( k \)th pathway for the \( i \)th patient, \( x_i^{(k)} = \{ x_{i1}^{(k)}, \ldots, x_{ip_k}^{(k)} \} \) be the vector of the genomic measures of the genes in the \( k \)th pathway for the \( i \)th patient, and let \( x_i = (x_i^{(1)}, \ldots, x_i^{(K)}) \) 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 \( X = \{X^{(1)}, \ldots, X^{(K)}\} \) 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,

\[
F(X) = \sum_{k=1}^{K} F_k(X^{(k)}),
\]

where \( F_k(X^{(k)}) \) can be interpreted as the activity level associated with the \( k \)th 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 \( Y \),

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

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 \( \gamma \). For the censored survival phenotype, we can assume that the hazard function at time \( t \) given the observed genomic data \( X \) is modeled as

\[
\lambda(t|X, Z) = \lambda_0(t) \exp(F(X) + \gamma Z),
\]

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

The main motivation for these models is that the known biological pathways naturally divide genes into sets and 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 “NPR” to particularly emphasize this point, i.e. the genetic and pathway effects are modeled nonparametrically. By doing so, we limit the genetic interactions only to within pathways and, therefore, greatly reduce the hypothesis space, enabling identification of pathways perturbed in the disease process. It is obvious that without any constraints on the functions \( F_k(X^{(k)}) \), Model (2.2) or (2.3) is not identifiable. In Section 3, we propose a general pathway-based GDB procedure to identify such NPR models with the particular form of (2.2) or (2.3).

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 (2.1) using regression trees (Breiman and others, 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 and others, 2000; 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 boosting algorithms can be seen as functional gradient descent techniques (Friedman, 2001). The goal is to estimate the function $F: R^p \rightarrow R$, minimizing an expected loss function

$$E[L(Y, F(X))], \quad L(\cdot, \cdot): R \times R \rightarrow R^+$$

based on data $(y_i, x_i), i = 1, \ldots, n$. Estimation of such a $F(\cdot)$ from data can be done via a constrained minimization of the empirical loss

$$n^{-1} \sum_{i=1}^{n} l(y_i, F(x_i))$$

by functional gradient descent (Friedman, 2001). The key step of such a functional gradient descent procedure is to project the negative gradient vector of the loss function evaluated at the sample values, denoted by $\tilde{y}_i, i = 1, \ldots, n$, to base learner, where the base learner can be viewed as an estimate of $E[\tilde{y}_i | x]$. The final estimate of the function $F(X)$ is then a linear combination of these base learners, where the coefficients are determined by minimizing the loss function during each boosting step.

There are many possibilities for choosing a base learner. Examples include the regression tree, linear model, and the component-wise smoothing spline (Bühlmann and Yu, 2003), of which the regression tree is the most popular choice. In general, to prevent overfitting, the base learner in boosting should be simple, involve only a few parameters, and have low variance relative to bias. 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 GDB over genes in each of the $K$ pathways separately and choose the pathway that gives the best fit to the negative gradient vector.

We first consider the case when no other covariates are included in Model (2.2) or (2.3). Let $L(y_i, F(x_i))$ be a loss function for the $i$th observation, which depends on the type of the phenotype. For binary phenotype and Model (2.2), the loss function can be defined as

$$L(y, F(x)) = \sum_{i=1}^{n} L(y_i, F(x_i)) = \sum_{i=1}^{n} \log(1 + \exp(-2y_i F(x_i))), \quad y_i \in \{-1, 1\}.$$  \hspace{1cm} (3.1)

This is also the loss function used by Friedman and others (2000) 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 (2.3),

$$L(y, \delta, F(x)) = \sum_{i=1}^{n} L(y_i, \delta_i, F(x_i))$$

$$= - \sum_{i=1}^{n} \delta_i \left\{ F(x_i) - \log \left( \sum_{j=1}^{n} 1_{\{y_j \geq y_i\}} \exp(F(x_j)) \right) \right\}.$$ \hspace{1cm} (3.2)

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, \( F^{(0)}(X) = 0, F_k^{(0)}(X^{(k)}) = 0, k = 1, \ldots, K \).

   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,

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

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

\[
(a^{(k)}, \beta^{(k)}) = \arg\min_{a, \beta} \sum_{i=1}^N [\tilde{y}_i - \beta h_k(x_i^{(k)}; a)]^2, \quad k = 1, \ldots, K,
\]

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

4. Line search over \( \rho \) for the pathway \( k^* \) selected in Step 3,

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

5. Updating the function with \( \nu \) being the learning rate,

\[
F^{(m)}_k(X^{(k)}) = F^{(m-1)}_k + \nu \rho_m h_{k^*}(x_i^{(k^*)}; a^{(k^*)}),
\]

\[
F^{(m)}(X) = F^{(m-1)}(X) + F^{(m)}_{k^*}(X^{(k^*)}).
\]

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 \( F^{(m)}_k(x^{(k)}) \) denotes the function \( F_k(x^{(k)}) \) at the \( m \)th 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

\[
\tilde{y}_i = -\left[ \frac{\partial L(y_i, F(x_i))}{\partial F_k(x_i^{(k)})} \right]_{F(X) = F^{(m-1)}(X)} = 2y_i/(1 + \exp(2y_i F^{(m-1)}(x_i))) \tag{3.3}
\]

for the logistic model (2.2) and

\[
\tilde{y}_i = \delta_i - \sum_{j=1}^n \delta_j 1_{[t_i \geq t_j]} \frac{\exp(F^{(m-1)}(x_i))}{\sum_{r=1}^n 1_{[t_r \geq t_j]} \exp(F^{(m-1)}(x_r))}
\]

for the Cox model (2.3). 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 Steps 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.

It should be emphasized that our proposed NPR models do not attempt to incorporate the pathway structures; rather, they use the pathway information to divide genes into sets. However, our models allow
interactions of genes within a given pathway. In order to model such interactions between genes in a given pathway, we propose to use a $J$-terminal node regression tree (Breiman and others, 1984) as the base learning procedure in Step 3 of the algorithm for each pathway. Regression trees provide a natural nonparametric function for modeling higher-order interactions between genes. For pathway $k$, each regression tree itself has the additive form

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

where $\{R^{(k)}_j\}, j = 1, \ldots, J$, are disjoint regions that cover the space of all joint values of the variables $X^{(k)}$, and $a^{(k)} = \{b^{(k)}_j, R^{(k)}_j\}_{j=1}^J$ in the general boosting algorithm for the NPR models (Step 4), where $b^{(k)}$ is the corresponding coefficients. Here $J$ controls the size of the tree and also the highest order of interactions. Note that the regression tree models interactions among the variables by imposing that a node depends on its parent nodes. The boosting procedure with regression trees as base procedures inherits the favorable characteristics of trees such as robustness, and flexibility in modeling interactions (Breiman and others, 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 and others, 1984).

Finally, if covariates $Z$ are included in the NPR models (2.3) or (2.2), we can iterate between updating the parametric parameters $\gamma$ 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 (2.2) 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 and others, 1984), the final function $F(X)$ identified by the pathways-based GDB procedure is a linear combination of trees and must therefore be interpreted in a different way. For a single tree $T$, Breiman and others (1984) proposed to use

$$I^2_l(T) = \sum_{t=1}^{J-1} \hat{i}^2_l I(v(t) = l)$$  \hspace{1cm} (3.4)

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 $t$th node. This score is basically the summation of the empirical improvement $\hat{i}^2_l$ 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 $l$th variable as

$$I^2_l = \frac{1}{M} \sum_{m=1}^M I^2_l(T_m),$$  \hspace{1cm} (3.5)

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

The importance scores defined by (3.4) and (3.5) 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 $k$th pathway, i.e.

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

where $M_k$ is the number of times the $k$th pathway was selected in Step 3 of the proposed pathway-based boosting algorithm, and $T_{mk}$ is the $m$th tree built based on the $k$th 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 $X_{(k-1)\times10+1}, \ldots, X_{(k-1)\times10+10}$ belong to the $k$th pathway, for $k = 1, \ldots, 5$. We generate $X_i$s independently from Bernoulli distribution with probability of 0.25 being 1. We generate disease status variable $Y$ based on the following logistic regression model:

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

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

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

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 learner 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 2 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 3 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, where for each model the number of top scoring variables is the true number of variables that are related to the risk of disease. 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 four SNPs in 59 and 58% of the simulations, respectively. The SNP1 appeared among the top three SNPs in 84% of the replications for Model 4. In addition, for Model 4, SNP2 and SNP12 appeared among the top three 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 4 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

![Fig. 2. Simulation results: frequencies of the pathways selected during the pathway-based boosting procedure over 100 replications for each of the four simulated models.](https://academic.oup.com/biostatistics/article-abstract/8/2/265/230606)
Fig. 3. Simulation results: the boxplots of the relative importance scores for the four pathways based on 100 replications for each of the four simulated models.

Table 1. Simulation results: percentage of the true variables that are among the genes with highest top two scores (for Model 1 and 2 where two variables X1 and X2 are related to the risk of disease), top four scores (for Model 3 where four variables (X1, X2, X11, X12) are related to the risk of disease), and top three scores (for Model 4 where three variables (X1, X2, X12) are related to the risk of disease) over 100 replications. NPR: proposed pathway-based boosting procedure for the NPR models; GDB: the gradient boosting procedure of Friedman

<table>
<thead>
<tr>
<th>Method</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X1</td>
<td>X2</td>
<td>X1</td>
<td>X2</td>
</tr>
<tr>
<td>NPR</td>
<td>0.81</td>
<td>0.82</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>GDB</td>
<td>0.66</td>
<td>0.68</td>
<td>0.41</td>
<td>0.47</td>
</tr>
<tr>
<td>SVM</td>
<td>0.90</td>
<td>0.82</td>
<td>0.32</td>
<td>0.42</td>
</tr>
</tbody>
</table>

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 SNP1 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 GDB procedure of Friedman and the popular SVM 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 using univariate variable selection 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. We should expect that the NPR method works better in the scenario when multiple pathways independently affect the disease risk (e.g. Models 3 and 4), in which case the univariate variable selection fails to identify the relevant variables.
4.4 Simulations based on more complex disease models

To further demonstrate the utility of the NPR model and the boosting algorithm, we simulated data from two complex disease models, denoted by Models 5 and 6. We assume that there are 50 candidate pathways, each having 10 SNPs with minor allele frequencies of 25% for a total of 500 SNPs. We denote the 500 candidate SNPs as $X_1, \ldots, X_{500}$, where SNPs $X_{(k-1)\times 10+1}, \ldots, X_{(k-1)\times 10+10}$ belong to the $k$th pathway, for $k = 1, \ldots, 50$. We simulate a baseline probability of getting disease at 12%. We assume that pathways 1, 2, and 3 can independently increase the risk of disease based on different mechanisms. For Model 5, we assume that for pathway 1, the probability of disease increases by

$$0.25(X_1X_2 \text{ OR } X_3X_4),$$

i.e. if there are mutations in both $X_1$ and $X_2$ or $X_3$ and $X_4$, then the probability of disease is increased by 0.25; for pathway 2, the probability of disease increases by

$$0.12(X_{11} + X_{12} + X_{11}X_{12});$$

and for pathway 3, the probability of disease is increased by

$$0.2 \times I[M \geq 4] + 0.2 \times I[M \geq 5] + 0.3 \times I[M \geq 6] + 0.2 \times I[M \geq 7] + 0.06 \times I[M \geq 9],$$

where $M = \sum_{j=1}^{10} X_{30+j}$ is the number of mutations in 10 SNPs in pathway 3. This type of “additive” model for pathway 3 was considered in Huang and others (2004) in their development of the FlexTree procedure. However, the overall disease model we consider here is more complex than that considered by Huang and others since we assume that three independent pathways can all lead to increased risk of disease. This model will result in a disease probability of around 25%. Model 6 is similar to Model 5 except that for pathway 1 we assume that the probability of disease increases by

$$0.50(X_1X_2 \text{ OR } X_3X_4),$$

i.e. the risk of disease due to pathway 1 is higher. For this model, the population disease rate is about 30%.

We simulate data sets of 500 individuals and for each model, we repeat the simulation 100 times. The top panel of Figure 5 shows the boxplots of the pathway relative importance scores for Model 5 and Model 6. Overall, we observe that the relevant pathways have higher relative importance scores than those irrelevant pathways. For Model 5, pathway 3 tends to have the highest relative importance scores; and for Model 6, pathway 1 tends to have the highest relative scores. For Model 5, among the 100 replications, pathways 1, 2, and 3 are among the top three pathways with the highest scores in 27, 60, and 79% of the simulations, respectively. As a comparison, for Model 6, among the 100 replications, pathways 1, 2, and 3 are among the top three pathways with the highest scores in 93, 42, and 75% of the simulations, respectively. This is expected since pathway 1 has a relatively weak effect on disease risk in Model 5 but a relatively strong effect in Model 6.

To examine how sample sizes affect the results, we simulate data sets of 1000 individuals, and again for each model, we repeat the simulation 100 times. The bottom panel of Figure 5 shows the boxplots of the pathway relative importance scores for Model 5 and Model 6. As expected, as sample size increases, more irrelevant pathways have relative importance scores of zero, indicating that these irrelevant pathways are less frequently selected during the boosting iterations. As a comparison to a sample size of 500, for Model 5, among the 100 replications, pathways 1, 2, and 3 are among the top three pathways with the highest scores in 43, 98, and 85% of the simulations, respectively; and for Model 6, among the 100 replications, pathways 1, 2, and 3 are among the top three pathways with the highest scores in 98, 89, and 79% of the simulations, respectively.
5. Applications to Real Data Sets

We present results of application of the NPR methods to three published breast cancer gene expression data sets, including the data sets of Wang and others (2005), Miller and others (2005), and Sotiriou and others (2006).

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

Wang and others (2005) reported large Affymetrix-based gene expression profiling for 286 patients with lymph-node-negative primary breast cancer. These patients were treated between 1980 and 1995 with age at surgery ranging 26–86 years and a median age at surgery of 52 years. No patient received any adjuvant therapy. During the follow-up period, 180 of these patients were relapse-free at 5 years, 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) to 81 genes (e.g. regulation of cell cycle).
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.

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

We first performed the analysis using the logistic regression model (2.2). 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 top left plot of Figure 6 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
Fig. 6. Results from analysis of breast cancer data sets: plot of the frequencies of the pathways selected and the pathway importance scores during the boosting procedure for the logistic regression models (left panel) and the Cox models (right panel) for the data sets of Wang and others (top panel) and Miller and others (bottom panel).

procedures are shown in Table 3. It is clear that the NPR outperforms almost all 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 (2.3) with time to cancer relapse as the outcome. The top right plot of Figure 6 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.

5.2 Application to two other breast cancer gene expression data sets

Miller and others (2005) reported a gene expression profiling study of 251 primary breast cancer tissues resected in Uppsala County, Sweden from January 1, 1987, to December 31, 1989, using Affymetrix Chip
Table 3. Comparison of the average misclassification error rates of the NPR method and nine commonly used procedures based on 10-fold cross-validation for three real data sets

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Data set</th>
<th>Wang and others</th>
<th>Miller and others</th>
<th>Sotiriou and others</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPR</td>
<td></td>
<td>0.29</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>Random forest</td>
<td></td>
<td>0.33</td>
<td>0.23</td>
<td>0.25</td>
</tr>
<tr>
<td>Decision stump</td>
<td></td>
<td>0.42</td>
<td>0.25</td>
<td>0.24</td>
</tr>
<tr>
<td>Logistic regression</td>
<td></td>
<td>0.36</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>Neural network</td>
<td></td>
<td>0.29</td>
<td>0.36</td>
<td>0.37</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td></td>
<td>0.34</td>
<td>0.30</td>
<td>0.24</td>
</tr>
<tr>
<td>Bayes net</td>
<td></td>
<td>0.40</td>
<td>0.26</td>
<td>0.23</td>
</tr>
<tr>
<td>Bagging</td>
<td></td>
<td>0.35</td>
<td>0.25</td>
<td>0.26</td>
</tr>
<tr>
<td>Ada boost(M1)</td>
<td></td>
<td>0.33</td>
<td>0.30</td>
<td>0.32</td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td>0.42</td>
<td>0.26</td>
<td>0.31</td>
</tr>
</tbody>
</table>

HG-133A and HG-133B (GEO Accession No. GSE3494). The authors identified an expression signature for p53 which can be used for predicting the mutation status, transcriptional effects, and patient survival. Among these patients, 236 of them had follow-up information in terms of time and event of disease-specific survival. The same 245 genes in 33 cancer-related sub-pathways used in the previous example (see Table 2 for the pathways and the number of genes in each pathway) were used in our analysis of this data set.

The plots in the bottom panel of Figure 6 show the relative importance scores of the pathways based on a logistic regression model (left plot) and the Cox regression model (right plot), indicating that the pathways related to Metalloendopeptidases (MMPs) and MMP inhibitors, as well as cell proliferation, cell growth, and maintenance are important to breast cancer-specific survival. The misclassifications for 5-year death due to cancer that results from various procedures based on 10-fold cross-validation are shown in Table 3, indicating that NPR outperforms almost all the competitors.

We also applied the NPR method to another breast cancer gene expression data set as reported in Sotiriou and others (2006) (GEO Accession No. GSE2990), including gene expression data from 189 invasive breast carcinomas. Among these 189 patients, 88 of them are from the data set of Miller and others (2005), and 101 are patients from the John Radcliffe Hospital (Oxford, UK). Treating the relapse-free survival time as the outcome, the NPR model with the Cox model identified the MMP pathway and the cell growth and maintenance pathway as the most important pathways related to breast cancer relapse. These two pathways are also the important pathways identified based on Miller’s data set for breast cancer-specific survival. These top scoring pathways are different from those identified from Wang’s breast cancer data set. One reason might be that Wang’s data set includes only lymph-node-negative primary breast cancer patients. The misclassifications for 5-year cancer relapse that results from various procedures based on 10-fold cross-validation are shown in Table 3, again indicating that NPR outperforms almost all the competitors.

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 genomic data, especially in identifying genes and their interactions and pathways that might be related to
the phenotypes. In this paper, we have introduced the NPR models and proposed to extend the gradient boosting procedure of Friedman (2001) to obtain fits of such models. The model has a natural interpretation as pathway activities. 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 three breast cancer data sets. Different from the traditional regression analysis, the proposed methods naturally incorporate biological pathways information. Different, also, from the commonly used GSEA, our method considers multiple pathways simultaneously.

It is important to note that in our proposed NPR models, the pathway information is only used to divide the genes into sets, and genes within pathways are allowed to interact to affect the pathway activities. However, the NPR models do not attempt to model the actual pathway structures. To the best of our knowledge, we have not seen any published work trying to statistically model the pathway/network structures in the context of relating genomic data to phenotypes. Besides the difficulty of statistically capturing the complex pathway structures, another reason is that our knowledge of the real pathway/network structure is still very limited. Pathway analysis using comprehensive knowledge of mammalian biology can greatly reduce the hypothesis space, enabling identification of pathways perturbed in the disease process. By utilizing the information which genes belonging to which pathways and by only considering genes that belong to sets of possible pathways, we can limit the number of genes to be considered and also the gene–gene interactions that need to be considered in modeling high-dimensional genomic data. Since the pathway-based GDB procedure chooses only one pathway during each boosting step, the algorithm should not have computational difficulty in handling hundreds of pathways simultaneously. The algorithm is applicable to typical pathway-based candidate gene association studies of complex diseases and also to genome-wide expression profiling data when combined with biological pathways information, as shown in our analysis of three breast cancer gene expression profiling data sets.

A related 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 misspecification 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 pathway activities affect the phenotype in an additive model conditioned on the genes that the pathways include. We are currently extending the NPR models to include pathway–pathway interactions defined by their respective pathway activity functions.

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. Other important future research includes thorough
evaluation of the pathway-based boosting procedure for continuous and censored survival phenotypes and for models with environmental covariates.

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. The R codes for implementing the proposed methods as well as detailed documentation can be downloaded from the authors’ website (http://www.cceb.upenn.edu/~hli/NPR).

ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health grant ES009911 and a grant from the Pennsylvania Department of Health. We thank Mr. Edmund Weisberg, MS at Penn Center for Clinical Epidemiology and Biostatistics for editorial assistance. We also thank two referees for their helpful comments. Conflicts of Interest: None declared.

REFERENCES


[Received February 9, 2006; first revision April 19, 2006; second revision June 9, 2006; accepted for publication June 12, 2006]