University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2005 *Paper* 197

Data Adaptive Pathway Testing

Merrill D. Birkner* Alan E. Hubbard†
Mark J. van der Laan‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, mbirkner@berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley, hubbard@stat.berkeley.edu

‡Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper197

Data Adaptive Pathway Testing

Merrill D. Birkner, Alan E. Hubbard, and Mark J. van der Laan

Abstract

A majority of diseases are caused by a combination of factors, for example, composite genetic mutation profiles have been found in many cases to predict a deleterious outcome. There are several statistical techniques that have been used to analyze these types of biological data. This article implements a general strategy which uses data adaptive regression methods to build a specific pathway model, thus predicting a disease outcome by a combination of biological factors and assesses the significance of this model, or pathway, by using a permutation based null distribution. We also provide several simulation comparisons with other techniques. In addition, this method is applied in several different ways to an HIV-1 dataset in order to assess the potential biological pathways in the data.

1 Introduction

1.1 Motivation

A majority of diseases are caused by a combination of factors, for example, composite genetic mutation profiles have been shown in many cases to predict disease. This is especially relevant to genetically caused diseases, where one mutation alone does not cause a deleterious state, but instead several mutations in combination will cause a specific disease outcome. Specific cancers, for example, may be influenced by the accumulation of and interactions between genetic mutations. Therefore determining these specific genetic models is vital when predicting the cancerous state of an individual.

Scientists are posed with the problem of determining the specific combination of factors which predicts an outcome, and the respective overall significance of a model. In order to predict a specific outcome based on a combination of variables, several methods have been proposed in the statistical literature. This article will outline an existing approach which will determine the significance of the association of a pathway (subset of biologically relevant factors) with an outcome. This technique will be compared to other methods in simulations. Additionally, the pathway procedure will be applied to an HIV-1 dataset. The method will be used to predict the viral replication from sets of codons.

1.2 Current Statistical Approaches

Several methods have been used to test biological pathways. This section will briefly outline the current procedures. Firstly, a specific pathway technique was proposed by Jelle Goeman to model these biological processes. This procedure is implemented in R and is referred to as globaltest(). The goal of this procedure is to test if a group of factors is associated with a given outcome. Therefore, in a situation with M genes, one is interested in testing a subgroup of these genes. The test will give one significance level, or p-value for each group of genes. One can therefore determine if a gene expression pattern is related to a specific clinical outcome (continuous or categorical). This approach is based on the following model setup: $E(Y_i|r_i) = h^{-1}(\alpha + r_i)$, where $r_i = \sum_j x_{ij}\beta_j$ for $i = 1, ..., n$. They describe this model as a random effects model in which each individual influences the outcome. The vector $r = r_1, ...r_n$ has $E(r) = 0$ and $Cov(r) = \tau^2 XX^T$. In order to test the null

hypothesis, $\tau^2 = 0$, a score test statistic is calculated and compared to a cX^2 distribution with a specific scaling factor, c. When the user is performing this pathway test on a small sample, there is an option in the function to implement the permutation method to calculate a respective p-value. The global pathway technique has also been extended to the case of a survival outcome (Goeman et al., 2005). In this case the clinical outcome is the survival of a patient. Again one is interested in determining subgroups, or pathways, of genes that predict survival.

In addition to the global test procedure, researchers can also use univariate tests of the variables in the model or pathway to test the relative significance of the entire pathway. These tests are then adjusted for multiple testing, since one or more repetition implies repetition of the null hypothesis of interest. This technique involves initially creating univariate tests of each gene in a subgroup. A multiple testing approach controlling family wise error (FWER) is then applied to the group of tests. A pathway is in turn declared significant if at least one of the variables has a adjusted p-value, adjusting for multiple testing, less than a pre-specified α level. This procedure does not inherently include interaction terms among the various genes.

The article will focus on a method which constructs a permutation null distribution for the test statistic of the model. As expanded upon in the subsequent sections, a data adaptive technique will be used to build the model. A similar method consists of performing a linear regression with the biological factors. This therefore assumes that these variables ($X = X_1, ...X_p$) are main effect terms, and does not incorporate the inclusion of a biological interaction between two or more factors. In addition, this method does not allow for a non-linear function of an individual factor. Such an approach is adequate if the model does not contain an interaction or if it is of a linear form, but models deviating from this form are not correctly modelled with this technique. Thus, this approach could lack power to detect a significant association if the true model deviates significantly from linear.

A special case of the method described in this paper is included as an option in the logic regression R function to test the significance of the model against the null model (Ruckinski et al., 2003). The logic regression method uses a data adaptive regression technique to build Boolean combinations of binary covariates. This regression technique allows main effect terms as well as interactions to be included in the model when predicting an outcome. The logic regression method includes an option referred to as the "null model test" (Ruckinski et al., 2003). This is a test for signal in the data and

compares the optimal logic model to the null model, which is based on the permutation distribution. The null hypothesis which is tested corresponds to independence between the dependent and independent variables. The model is built on a dataset corresponding to the original covariates and a permuted set of outcome variables. Each model produces a score and the proportion of model scores under this null model which exceed the optimal model corresponds to the p-value of the optimal model. This is implemented in the R function logreg() in the LogicReg library (Ruckinski et al., 2003). To the best of our knowledge the application of the permutation based null distribution and data adaptive regression is only used in Logic Regression, and has not been widely advertised as a method to test the significance of a model built by a data adaptive regression technique. Our purpose is to generalize this approach and add a formal argument to why this approach gives exact inference.

2 Methods

2.1 Data Adaptive Approach

We denote the data as (Y_i, X_i), where $i = 1, ..., n$. The outcome measure is Y_i and the covariates of interest is a vector X_i for each individual i. We wish to test the null hypothesis: $H_0 : X \perp Y$. In order to determine the optimal model which best predicts the outcome from a group of variables, a data adaptive regression technique will be implemented. Data adaptive regression procedures use selection criteria to choose the variables and forms of variables to place in the model, including in some cases, linear spline terms. The pathway method outlined in this article can use any data adaptive procedure, but for the sake of the simulation example, we illustrate the **polyclass()** method in R (Kooperberg et al., 1997).

POLYCLASS is an exploratory, data-adaptive, or black box regression technique used to predict categorical or binary outcomes. This classification method, uses forward addition and backward deletion, searches through a series of models defined by main effects, splines and cross-products to create a logistic regression model. The procedure uses cross-validation to choose the complexity (number of basis functions) of the model. With respect to the addition steps, proposed new predictors are either main effects not already in the model; knots to existing main effects creating linear spline terms or;

any interaction terms already in the model. For the deletion step, terms are removed hierarchically (e.g., a main effect term is not removed before it's corresponding interaction term).

An alternative method to the polyclass() function is the deletion, substitution, and addition algorithm, proposed by Sinisi and van der Laan (2004). The Cross-validated Deletion/Substitution/Addition (D/S/A) algorithm (Sinisi and van der Laan, 2004) is a data-adaptive machine learning methodology, which is used to predict an outcome or response, Y, given a set of covariates. The algorithm minimizes the expectation of a specific loss function. In the case of a continuous outcome, the loss function is the residual sum of squares and the parameter of interest is the expectation of the outcome given the covariates. The algorithm is based on deletion, substitution, and addition moves which build models of varying dimensions. The final model is chosen as the model that minimizes the cross validated residual sum of squares. Finally, other data adaptive techniques include logic regression or MARS (Ruckinski et al., 2003; Friedman, 1991).

2.2 Summary Statistic

Once a model is built using a data adaptive regression technique a summary statistic of this model is computed. This summary statistic compares the built model to the model including only the intercept. In the case of a logistic regression model, we will use the likelihood ratio statistic, $2*log(LR_0 - LR_1)$. The tail probability (p-value) of this summary statistic is then computed under a X^2 distribution with the degrees of freedom equaling the difference in the number of parameters between the full and null model. We will estimate the null distribution of the p-value using a permutation approach.

2.3 Permutation Null Distribution

After obtaining the reference p-value from the summary statistic of the proposed model, the relative significance of this p-value in comparison to the null distribution must be calculated. The method will revolve around the construction of the null distribution, for which the null hypothesis, $H_0 : Y \perp X$, holds. The permutation distribution can be shown to provide correct error control under the null hypothesis in the following manner: If the null hypothesis is true, and therefore $Y \perp X$, then the conditional distribution of the data given the marginal empirical distributions of $(X_1, ..., X_n), (Y_1, ..., Y_n)$

of X and Y, respectively, equals the permutation distribution. Therefore controlling the Type-I error under the permutation distribution corresponds with controlling the Type I error under the true conditional distribution if H_0 is true. In particular, this shows that if H_0 is true, then the Type I error control under the permutation distribution implies Type I error control under the true distribution.

In order to create this null distribution, the outcome values are permuted, therefore corresponding to the situation where there is no association between the dependent and independent variables. The data is permuted and the **polyclass()** function is applied to the data and a model is built using the same data adaptive technique. The p-value from the likelihood ratio summary statistic is computed, by comparing it to a specific X^2 distribution. This permutation method is repeated for each unique permutation of the data (or a practical number of times, e.g. 5000) and a distribution of p-values is calculated. The p-value of the overall association of the outcome and the risk factors is simply the survival function of the observed p-value in the original experimental dataset, with respect to the permutation distribution: The pathway p-value$= S_{perm}(\hat{p}) \equiv P(X \geq \hat{p})$, where X's distribution under the null is the permutation distribution.

3 Simulations

3.1 Type I error rate

Before applying this method to biological datasets, we must determine the Type-I error control of this method. In order to determine this control, we simulated data in which the outcome was independent of the covariates. We simulated an outcome from a Binomial distribution with probability equal to 0.5. There were $M = 20$ covariates, $X = X_1, ...X_{20}$, which were simulated from a Multivariate Normal distribution with mean 0 and correlation matrix of 0.8 between all variables, and 1 between the diagonal elements. The sample sizes studied included $n = 200, 500, 700$.

We simulated 500 datasets, each using 3000 unique permutations for the creation of the null distribution; the null distribution was the permutation data null distribution.

Table 1 indicates that this method has the correct Type-I error control (the probability of rejecting the null is at the specified α-level). This therefore

Table 1: Type-I Error Control.

Simulation Parameters	Type-I error (out of 500 iterations)
$Y \sim B(p = 0.5), X \sim mvn(0, \Sigma = 0.8), n = 200, M = 20$	0.044
$Y \sim B(p = 0.5), X \sim mvn(0, \Sigma = 0.8), n = 500, M = 20$	0.052
$Y \sim B(p = 0.5), X \sim mvn(0, \Sigma = 0.8), n = 700, M = 20$	0.048

indicates that the procedure has correct error control in correctly accepting models in which the Y is independent of the X covariates.

Testing methods attempt to control Type I error rate while simultaneously maximizing power of the procedure. This previous section indicated that this procedure seems to have proper control over 500 simulated datasets.

Simulations are now presented which were performed to determine the power of the procedure in various scenarios. The simulations included a binary outcome Y (derived from an underlying model), $n = 500$, and $\epsilon_i \sim N(0, 1)$ error terms was added to the linear term when constructing the probability. Therefore, when p is equal to the number of covariates in the model, the probability can be written as:

$$P(Y = 1|X) = \frac{e^{(\beta_0+\beta_1 X_1+...\beta_p X_p+\epsilon)}}{1 + e^{(\beta_0+\beta_1 X_1+...\beta_p X_p+\epsilon)}}. \quad (1)$$

The first dataset was created with $M = 10$ terms, or covariates, and and underlying model of $logit(P) = \beta_0 + \beta_1 X_1 X_2 + \epsilon$, with $\beta_0 = 5.179$, and $\beta_1 = 5.154$. Therefore, the model used to construct the probability of $Y = 1$ was:

$$P(Y = 1|X) = \frac{e^{(\beta_0+\beta_1 X_1 X_2+\epsilon)}}{1 + e^{(\beta_0+\beta_1 X_1 X_2+\epsilon)}}. \quad (2)$$

In addition to the polyclass pathway technique, the regression method with main effects, Golubtest, and the FWER technique were compared. Both the Polyclass method as well as the Logistic Regression, Backward and Forward and Backward regression used the permutation distribution. The Regression $X_1 + X_2$ method consists of testing a logistic regression with two

Table 2: Comparison of Methods

Method	Power (rejections/500 iterations)
Polyclass	0.762
Logistic Regression, $X_1 + X_2$	0.204
Golubtest (Pathway 1:2)	0.18
Logistic Regression, 10 main terms	0.12
Logistic Forward and Backward Selection	0.15
Logistic Backward Selection	0.10
FWER	0.062

main effect terms. Golubtest (Pathway 1:2) also used X_1 and X_2. The regression with 10 main terms tested a logistic regression with each of the 10 Xs as a main effect term. The forward and backward selection and forward selection methods were separately applied to the logistic model with 10 main effect terms. Finally, the FWER technique applied the Bonferroni adjustment to univariate tests of each of the 10 variables on the outcome. If any of these 10 adjusted p-values were less than 0.05 the pathway was claimed significant.

Table 2 indicates that given an interaction term in the model, the polyclass model is most efficient at picking up a model and claiming it significant. The models assuming either all main terms, a limited number of main terms, or various types of selection or multiple testing are not optimal, with respect to power, given the underlying model.

We next simulated data from an underlying model where the truth was X^2 where $\beta_0 = 1.179$, $\beta_1 = 2.154$, and $\epsilon_i \sim N(0, 1)$ (we put in an additional error term so that the true model is not a simple logistic regression). This can be written as follows:

$$P(Y = 1|X) = \frac{e^{(\beta_0+\beta_1 X_1^2+\epsilon)}}{1 + e^{(\beta_0+\beta_1 X_1^2+\epsilon)}}. \tag{3}$$

The Xs were drawn from either a normal or exponential distribution. When the X_1's are $N(0, 1)$ the resulting logit function over the observed X_1's is concave in X_1, when the X_1's are exponential ($\lambda = 1$), the function

Table 3: Truth is X_1^2:

Truth/Method	α	Power (rejections/500 iterations)
Polyclass (Normal)	0.05	0.220
Polyclass (Normal)	0.10	0.270
Polyclass (Exponential)	0.05	1
Polyclass (Exponential)	0.10	1
Logistic Regression (Normal)	0.05	0.01
Logistic Regression (Normal)	0.10	0.02
Logistic Regression (Exponential)	0.05	1
Logistic Regression (Exponential)	0.10	1

is nearly linear in X_1. The polyclass method was applied to the data to determine the respective power, as well as a method that builds a logistic regression with X_1 as its only main effect term.

These simulations (Table 3) indicate that in the case of exponential X values both the polyclass and logistic model pick up the desired model. The logistic model technique seems to be approximating the exponential shape with a line which is adequate in this situation. In the case where the Xs come from a normal distribution, the logistic model has a difficult time picking up this model, and polyclass does provide an improvement, although other black box regression techniques could do a better job of predicting this quadratic curve.

Finally, Table 4 compares the two methods when the truth was merely one X variable. Again, in the case of a regression, a logistic regression was fit with X_2 as its only main effect term. The underlying model is $logit(P) = \beta_0 + \beta_1 X_2$, where $\beta_0 = 0.179$, $\beta_1 = 0.154$, and the errors of the linear term were $\epsilon_i \sim N(0,1)$. The 10 X variables originated from $N(0,1)$ distribution.

In this case, the regression technique seems to outperform the polyclass technique. Both procedures are using the permutation null distribution, but with only one variable, the main effect regression technique could be producing a simpler model as compared to the polyclass technique, thus resulting in higher power.

Table 4: Truth is X_2:

Truth/Method	Power (rejections/500 iterations)
Polyclass	0.086
Logistic Regression	0.284

4 Data Analysis

Throughout this section we will refer to an application to an HIV-1 dataset. The described method was applied in a variety of ways to this dataset. The subsequent sections will initially outline the dataset and discuss the various applications of the pathway technique. All of these techniques resulted in findings which parallel previous statistical and biological findings.

4.1 HIV-1 Dataset

Studying the sequence of the Human Immunodeficiency Virus Type 1 (HIV-1) genome could potentially give important insight into the genotype–phenotype associations of the virus and in turn for the Acquired Immune Deficiency Syndrome (AIDS).

The phenotype which is studied in this dataset is the replication capacity (RC) of HIV-1, as it reflects the severity of the disease. A measure of replication capacity may be obtained by monitoring viral replication in an ideal environment, with many cellular targets, no exogenous or endogenous inhibitors, and no immune system responses against the virus (Barbour et al., 2002; Segal et al., 2004).

The genotypes correspond to codons in the protease and reverse transcriptase regions of the viral strand. The protease (PR) enzyme affects the reproductive cycle of the virus by breaking protein peptide bonds during viral replication. The reverse transcriptase (RT) enzyme synthesizes double-stranded DNA from the virus' single-stranded RNA genome, hence facilitating integration of the virus' genetic material into the host's chromosome. Since the PR and RT regions are essential to viral replication, many antiretrovirals (protease inhibitors and reverse transcriptase inhibitors) have been developed to target these specific genomic locations. Studying PR and RT genotypic variation involves sequencing the corresponding HIV-1 genome

regions and determining the amino acids encoded by each codon (i.e., a codon corresponds to a group of three nucleotides).

The HIV-1 sequence dataset consists of $n = 317$ records, linking viral replication capacity (RC) with protease (PR) and reverse transcriptase (RT) sequence data, from individuals participating in studies at the San Francisco General Hospital and Gladstone Institute of Virology (Segal et al., 2004). Protease codon positions $pr4 - pr99$ and reverse transcriptase codon positions $rt38 - rt223$ of the viral strand are studied in this analysis.

The outcome/phenotype of interest is the natural logarithm of a continuous measure of replication capacity, ranging from 0.261 to 151. The 282 covariates correspond to the codon positions in the PR and RT regions, with the number of possible codons ranging from one to ten at any given location. A majority of patients typically exhibit one amino acid at each position. Codons are therefore recoded as binary covariates, with value of zero corresponding to the most common amino acid among the $n = 317$ patients and value of one for all other amino acids, thus corresponding to a mutation. Previous biological research was used to confirm mutations and hence provide accurate PR and RT codon genotypes for each patient. The data for each of the $n = 317$ patients therefore consist of a replication capacity outcome/phenotype Y and an 282–dimensional covariate vector $X = (X(m) : m = 1, \ldots, M)$ of binary codon genotypes in the PR and RT HIV-1 regions.

4.2 Testing a Single Pathway

The initial data analysis which was applied to the HIV-1 dataset consisted of testing a subset of codons against the outcome of replication capacity. The subset of codons consisted of those mutations in the protease and reverse transcriptase positions which are known from the literature to be predictive of a change in replication of the virus. A detailed review of these positions can be found in Birkner et al. (2005). These positions consisted of: $rt184$, $rt215$, $rt41$, $rt210$, $rt116$, $rt65$, $rt67$, $rt69$, $rt70$, $pr54$, $pr53$, $pr46$, $pr47$, $pr48$, $pr50$, $pr36$, $pr77$, $pr82$, $pr32$, $pr84$, $pr20$, $pr30$, $pr24$, $pr73$, $pr88$, $pr10$, $pr90$, $pr93$, $pr71$, $pr63$.

In the applications with the HIV-1 dataset, the data adaptive method which was used to build the models was polymars() in the R library(polspline). This method is similar to the polyclass method with the exception that it is adapted to continuous outcomes, whereas the polyclass method is based on

the logit function and therefore adapted to binary or categorical outcomes. The function polymars is an adaptive regression procedure which uses linear splines to model the response. Therefore this method examines all main effects, interactions and splines to model the outcome by a set of predictor variables.

The pathway analysis was performed with the outcome $Y_i, i = 1, ..., n$ corresponding to the $n = 317$ replication capacity values. The $M = 30$ codon positions listed above corresponded to X matrix, which is of dimension $n \times M$. The polymars data adaptive regression function resulted in the following model corresponding to the prediction of replication capacity by a set of covariates:

$$ln(RC) = 3.598 - 3.261(pr32) - 0.542(rt184) + 0.369(pr47) \quad (4)$$

The summary statistic of this model was the F-statistic and a respective tail probability (p-value) was obtained by comparing this statistic to the F-distribution. This model was tested against the null hypothesis, in which the replication capacity is independent of the covariates. The n replication capacity values were permuted and the polymars() model was subsequently run. This was repeated 5000 times, each time a resulting p-value of the summary F-statistic was recorded. The corresponding p-value of the above model was compared to these 5000 p-values to determine its significance as compared to the null distribution. The final p-value for the above model was 0.0001, therefore indicating the significance of this model. Biological results of the individual codons in this model will be further discussed below.

4.3 Testing Multiple Pathways

In addition to testing a single pathway in the HIV-1 virus, it is also interesting to test the spatial significance of groups of codons. Therefore, in this example, one is only interested in the linear form of the virus, thus ignoring spatial interactions. Starting at the first codon, $pr4$, non-overlapping groups of 6 codons are tested separately against the outcome of replication capacity. In total there are 282 positions, and considering neighboring groups of 6 codons, there will be 47 groups. For example, the pathway method will be applied to $pr4$, $pr5$, $pr6$, $pr7$, and $pr8$, and a p-value will be recorded, which compares this model to the null distribution. The procedure is identical to

Table 5: 15 Most Significant Codon Groups

Codons	p-value
$pr10$, $pr11$, $pr12$, $pr13$, $pr14$, $pr15$	0.0001
$pr28$, $pr29$, $pr30$, $pr31$, $pr32$, $pr33$	0.0001
$pr34$, $pr35$, $pr36$, $pr37$, $pr38$, $pr39$	0.0001
$pr40$, $pr41$, $pr42$, $pr43$, $pr44$, $pr45$	0.0001
$pr46$, $pr47$, $pr48$, $pr49$, $pr50$, $pr51$	0.0001
$pr52$, $pr53$, $pr54$, $pr55$, $pr56$, $pr57$	0.0001
$pr70$, $pr71$, $pr72$, $pr73$, $pr74$, $pr75$	0.02
$pr82$, $pr83$, $pr84$, $pr85$, $pr86$, $pr87$	0.01
$pr88$, $pr89$, $pr90$, $pr91$, $pr92$, $pr93$	0.0001
$rt38$, $rt39$, $rt40$, $rt41$, $rt42$, $rt43$	0.01
$rt98$, $rt99$, $rt100$, $rt101$, $rt102$, $rt103$	0.045
$rt116$, $rt117$, $rt118$, $rt119$, $rt120$, $rt121$	0.0001
$rt134$, $rt135$, $rt136$, $rt137$, $rt138$, $rt139$	0.035
$rt182$, $rt183$, $rt184$, $rt185$, $rt186$, $rt187$	0.0001
$rt212$, $rt213$, $rt214$, $rt215$, $rt216$, $rt217$	0.01

the procedure described above in which a model is fit using polymars and the p-value corresponding to the summary F-statistic is recorded. This p-value is compared to a distribution of p-values obtained by running polymars on a permuted dataset 5000 times. Subsequently, $pr9$, $pr10$, $pr11$, $pr12$, and $pr13$ will be tested, and so on. This testing procedure will allow one to examine the spatial significance of the codons. The final 47 respective p-values will be plotted to examined.

The results of this analysis are presented in Figure 1 and Figure 3. Figure 3 illustrates a linear schematic of the virus and those blocks highlighted in pink correspond to groups of codons with a p-value less than 0.05 and those regions highlighted in blue correspond to areas with a p-value greater than 0.05. Figure 1 plots the p-value of each of these codon groups. Table 5 outlines the 15 codon groups with p-values less than 0.05.

In addition to testing neighboring groups of codons, this procedure was also applied to overlapping groups of codons. Therefore in this case, each codon was tested in a group with the two neighboring codons on either side. The groups of 5 codons are tested against replication capacity and the re-

spective p-value is recorded. The procedure is identical to the procedure described above in which a model is fit using polymars and the p-value corresponding to the summary F-statistic is recorded. This p-value is compared to a distribution of p-values obtained by running polymars on a permuted dataset 5000 times. This method results in a smoother distribution as compared to performing the pathway analysis on the disjoint sets of codons. The respective p-values are plotted in Figure 2. Again, the significant areas closely correspond to the areas highlighted in Table 5.

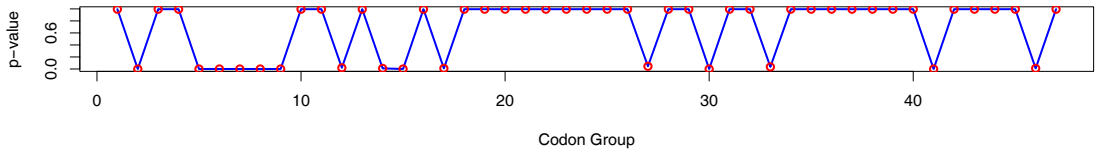

Figure 1: Neighboring Codon Groups.

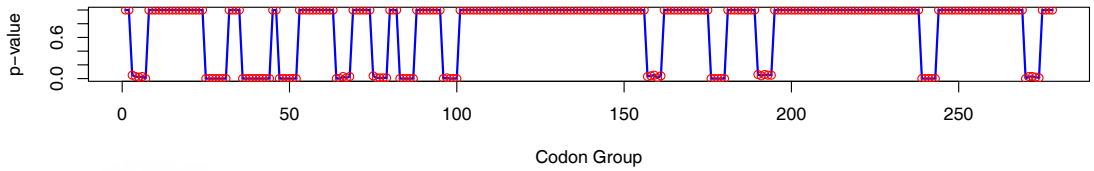

Figure 2: Overlapping Codon Groups.

4.3.1 Application of Multiple Testing

When comparing multiple pathways simultaneously, as illustrated above, one can apply a multiple testing procedure to the final list of marginal pathway p-values. Multiple testing procedures such as Benjamini and Hockberg's FDR procedure will control $E(\frac{V_n}{R_n}) \leq \alpha$, where V_n corresponds to the number of

Table 6: Regions of Viral Strand

Model	p-value
$ln(RC) = 3.517 - 1.218(pr43)$	0.0012
$ln(RC) = 3.551 - 0.2904(pr43) - 0.3039(pr46) - 2.046(pr43)(pr46)$	0.0014
$ln(RC) = 3.517 - 0.0869(pr47) - 0.4054(pr43) - 3.489(pr43)(pr47)$	0.00146

false positives and R_n refers to the number of rejections. In addition other marginal techniques can easily applied to the set of p-values.

4.4 Testing a Region of the Viral Strand

The pathway method can also be used to test sections of the viral strand. In this specific application the univariatly significant codons $pr43$, $pr46$, and $pr47$ were tested in successively larger groups with the data adaptive regression procedure polymars. Therefore three models were built with ($pr43$), ($pr43$, $pr46$), and ($pr43$, $pr46$, $pr47$) respectively. These positions were chosen because of their proximity on the viral strand as well as their significant univariate significance.

The polymars data adaptive regression technique was applied and a summary F-statistic was calculated, comparing the final model to the model containing only the intercept. The replication capacity values were permuted and this procedure was replicated 5000 times. The final model specific F-statistic was compared to the 5000 F-statistics under the null hypothesis ($H_0 : Y \perp X$). A final p-value is calculated which is recorded in Table 6.

Table 6 indicates that the three successively larger models of neighboring codons are significant when compared to the null model. The regression models indicate that an interaction between either $pr43$ and $pr46$ or $pr43$ and $pr47$ decreases the replication capacity to a greater extent as compared to a single codon mutation in position $pr43$, $pr47$ or $pr46$.

4.5 Testing Bivariate Models of Codons

In addition to testing neighboring sections of the viral strand, we were also interested in applying the pathway methodology to test bivariate models of

Table 7: Univariate Codon Significance

Codon	p-value
$pr10$	0.0002
$pr32$	0
$pr43$	0.0012
$pr44$	0.0001
$pr45$	0
$pr54$	0
$pr55$	0.0002
$pr60$	0.0355
$pr71$	0.001
$pr73$	0.024
$pr82$	0
$rt41$	0
$rt67$	0.0225
$rt83$	0.0145
$rt102$	0.003
$rt121$	0
$rt184$	0
$rt211$	0.0215
$rt214$	0.027
$rt215$	0.001

univariatly significant codons across the viral strand. A group of 20 codons was identified as those codons with marginal univariate p-values less than 0.05. These codons included: $pr10$, $pr32$, $pr43$, $pr46$, $pr47$, $pr54$, $pr55$, $pr60$, $pr71$, $pr73$, $pr82$, $rt41$, $rt67$, $rt83$, $rt102$, $rt121$, $rt184$, $rt211$, $rt214$, $rt215$. Initially a model was built for each of these positions to examine the respective individual univariate relationship with replication capacity. An F-statistic was obtained for the model. The replication capacity was permuted and the model was built 5000 times, each time recording a respective F-statistic. For each codon position, the respective F-statistic was compared to the null distribution of F-statistics and a corresponding p-value was calculated. Table 7 outlines these univariate p-values.

In addition to these univariate models, polymars was used on all 190

Table 8: Bivariate Models

Model	p-value
$ln(RC) = 3.528 - 0.2021(pr10) - 0.00086(rt102) - 2.268(pr10)(rt102)$	0.0001
$ln(RC) = 3.551 - 0.2904(pr43) - 0.3039(pr55) - 2.046(pr43)(pr55)$	0.0001
$ln(RC) = 3.512 - 0.3751(pr43) + 0.0483(rt102) - 3.121(pr43)(rt102)$	0.0002
$ln(RC) = 3.572 - 0.1753(pr43) - 0.2461(rt215) - 1.953(pr43)(rt215)$	0.0001
$ln(RC) = 3.549 - 4.013(pr55) - 0.4509(rt67) + 3.999(pr55)(rt67)$	0.0001
$ln(RC) = 3.497 - 0.5282(pr55) + 0.0399(rt102) - 2.823(pr55)(rt102)$	0.0001
$ln(RC) = 3.501 - 0.1039(pr71) + 0.1555(rt102) - 1.543(pr71)(rt102)$	0.0006

unique bivariate combinations of these 20 codons. As mentioned in the previous example, polymars was applied to two codons at a time and an F-statistic for the respective model was obtained. The replication capacity measurements were then permuted and the polymars procedure was repeated 5000 times, each repetition producing an F-statistic. Finally, the model F-statistic is compared to these 5000 null F-statistics and a respective p-value is obtained. This procedure is individually repeated for each of the 190 pairs of codons. We are particularly interested in the bivariate models in which the p-value is less than the minimum of the two p-values for the univariate models. This phenomena occurred seven times and the models and respective p-values are illustrated in the Table 8.

The models in Table 8 show the importance of an interactive effect of the codon mutations on the outcome of viral replication. As will be discussed in the following section, individually each of these codons is biologically important. It is therefore interesting to see the added effect produced by the combination of these mutations as compared to a single mutation alone. In particular, the fifth model containing $pr55$ and $rt67$ is interesting since the interaction of the two mutations causes an increase in viral replication. The other models correspond to cases where the combination of the two mutations decrease the replication capacity.

4.6 Biological Results

The above pathway analyses produced results that are biologically relevant to previous studies. In this section we will examine the individual importance of the positions mentioned in the previous sections. In particular, protease positions $pr32$, $pr34$, $pr43$, $pr46$, $pr47$, $pr54$, $pr55$, $pr82$, and $pr90$, and reverse transcriptase positions $rt41$, $rt184$, and $rt215$, have been singled out in previous research as related to replication capacity and/or antiretroviral resistance (Birkner et al., 2004; Segal et al., 2004; Shafer et al., 2001a). The specific mutations observed in our dataset parallel those found in the literature. For example, $Vpr32I$, $Mpr46I$, $Ipr54V/L/T$, $Vpr82A/T/F/S$, and $Lpr90M$, correspond to protease positions in which mutations increase the resistance to various protease inhibitors. Mutations in several of the identified codons also have an impact on the replication capacity of the virus. Reverse transcriptase mutation at position $rt41$ ($Mrt41L$) increases azidothymidine (AZT) resistance when present with $Trt215Y/F$. In addition, mutation $Mrt184V/I$ suppresses the wild-type activity of $Trt215Y$, thus decreasing AZT resistance (Shafer et al., 2001a). AZT, also known as Zidovudine, is a nucleoside reverse transcriptase inhibitor. It affects HIV's ability to replicate by producing faulty reverse transcriptase and hence inhibiting the transcription of RNA to DNA.

Several reverse transcriptase codon position mutations are related to antiretroviral resistance and viral replication capacity (i.e. positions $rt184$, $rt215$, $rt41$, $rt210$, $rt116$, $rt65$, $rt67$, and $rt69$) (Shafer et al., 2001b). Examples of such mutations include $rt41$, where $Mrt41L$ increases AZT resistance when present with a $Trt215Y/F$ mutation. A popular codon position, $Mrt184V/I$, partially suppresses the $Trt215Y$ mediated AZT resistance. Additionally, mutations at positions of the reverse transcriptase $rt41$, $rt184$, $rt215$ among others have shown resistance to NRTIs (Shafer et al., 2001b). Reverse transcriptase positions $rt215$, $rt184$, and $rt41$ have the largest resistance to AZT, as compared to other RT positions. Reverse transcriptase position $rt70$ mutations also causes resistance to AZT when there are amino acid changes at $Krt70R$, followed by $Trt215F/Y$, $Mrt41L$, $Drt67N$, and $Krt219Q$ (Goudsmit et al., 1997).

Finally, several protease mutations in certain positions have been found to have an impact on resistance of the virus (codons: $pr54$, $pr53$, $pr46$, $pr47$, $pr48$, $pr50$, $pr36$, $pr77$, $pr82$, $pr32$, $pr84$, $pr20$, $pr30$, $pr24$, $pr73$, $pr88$, $pr10$, $pr90$, $pr93$, $pr71$, $pr63$) (Shafer et al., 2001b). Protease positions $pr10$, $pr46$,

$pr48$, $pr54$, $pr63$, $pr71$, $pr82$, $pr84$, and $pr90$ cause resistance to saquinavir alone or in combination with AZT (Prado et al., 2002); protease positions $pr20$, $pr33$, $pr36$, $pr46$, $pr54$, $pr63$, $pr71$, $pr82$, $pr84$, and $pr90$ cause ritonavir resistance; protease positions $pr10$, $pr20$, $pr24$, $pr32$, $pr46$, $pr54$, $pr63$, $pr64$, $pr71$, $pr82$, $pr84$, and $pr90$ cause resistance to indinavir; and finally protease positions $pr30$, $pr36$, $pr46$, $pr71$, $pr77$, and $pr84$ cause resistance to nelfinavir.

5 Summary

This article considered a method to determine the significance of a biological pathway constructed with a data adaptive regression technique. This procedure uses the permutation distribution as the null distribution and it is shown how this distribution correctly controls the Type I error rate. Simulations were presented to verify the Type-I error control as well as compare this procedure to other currently used pathway analysis techniques. In addition, a data analysis has been presented, which applies the pathway technique in a variety of ways to the HIV-1 dataset. This procedure found potential new results when examining the various pathways. The results of this procedure are biologically interesting and parallel results found in the literature or from other statistical analyses. This procedure is an important method to be used in the field of computational biology to test the independence of genetic markers and an outcome.

We present a flexible procedure which can be used with both binary and continuous independent and dependent variables. It is important to note that, as discussed previously, the Logic Regression technique of Ruckinski et al. (2003) is testing the null hypothesis of independence between covariates and the outcome based on a logic regression fit. This approach is an appropriate method when one is dealing with binary covariates.

References

J. D. Barbour, T. Wrin, R. M. Grant, J. N. Martin, M. R. Segal, C. J. Petropoulos, and S. G. Deeks. Evolution of Phenotypic Drug Susceptibility and Viral Replication Capacity during Long-Term Virologic Failure of Protease Inhibitor Therapy in Human Immunodeficiency Virus-Infected Adults. *Journal of Virology*, 76(21):11104–11112, 2002.

M. D. Birkner, S. E. Sinisi, and M. J. van der Laan. Multiple testing and data adaptive regression: An application to HIV-1 sequence data. Technical Report 161, Division of Biostatistics, University of California, Berkeley, 2004. URL `www.bepress.com/ucbbiostat/paper161`.

M. D. Birkner, S. E. Sinisi, and M. J. van der Laan. Multiple testing and data adaptive regression: An application to HIV-1 sequence data. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. URL `http://www.bepress.com/sagmb/vol4/iss1/art8`.

J.H. Friedman. Multivariate Adaptive Regression Splines (with discussion). *The Annals of Statistics*, 19:1–141, 1991.

Jelle J. Goeman, Jan Oosting, Anne-Marie Clenton-Jansen, Jakob K. Anninga, and Hans C. van Houwelingen. Testing Association of a Pathway with Survival using Gene Expression Data. *Bioinformatics*, 21(9), 2005.

Jaap Goudsmit, Anthony de Ronde, Ester de Rooij, and Rob de Boer. Broad Spectrum of in Vivo Fitness of Human Immunodeficiency Virus Type 1 Subpopulations Differing at Reverse Transcriptase Codons 41 and 215. *Journal of Virology*, 71(6):4479–4484, 1997.

C. Kooperberg, S. Bose, and C. Stone. Polychotomous Regression. *Journal of the American Statistical Association*, 92:117–127, 1997.

Julia G. Prado, Terri Wrin, Jeff Beauchaine, Lidia Ruiz, Christos J. Petropoulos, Simon D.W. Frost, Bonaventura Clotet, Richard T. D'Aquila, and Javier Martinez-Picado. Amprenavir-Resistant HIV-1 Exhibits Lopinavir Cross-Resistance and Reduced Replication Capacity. *AIDS*, 16 (7):1009–1017, 2002.

Ingo Ruckinski, Charles Kooperberg, and Michael LeBlanc. Logic Regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, 2003.

M. R. Segal, J. D. Barbour, and R. M. Grant. Relating HIV-1 Sequence Variation to Replication Capacity via Trees and Forests. 3(1):Article 2, 2004. URL www.bepress.com/sagmb/vol3/iss1/art2.

R. W. Shafer, K. M. Dupnik, M. A. Winters, and S. H. Eshleman. A Guide to HIV-1 Reverse Transcriptase and Protease Sequencing for Drug Resistance Studies. In *HIV Sequencing Compendium*, pages 83–133. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, 2001a.

Robert W. Shafer, Kathryn M. Dupnik, Mark A. Winters, and Susan H. Eshleman. A Guide to HIV-1 Reverse Transcriptase and Protease Sequencing for Drug Resistance Studies. In *HIV Sequencing Compendium*, pages 83–133. Theoretical Biology and Biophysics Group at Los Alamos National Laboratory, 2001b.

Sandra E. Sinisi and Mark J. van der Laan. Deletion/Substitution/Addition Algorithm in Learning with Applications in Genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL http://www.bepress.com/sagmb/vol3/iss1/art18. Article 18.

Figure 3: Neighboring Codon Groups.