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

Year 2004 *Paper* 161

Multiple Testing and Data Adaptive Regression: An Application to HIV-1 Sequence Data

Merrill D. Birkner* Sandra E. Sinisi†
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

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

http://biostats.bepress.com/ucbbiostat/paper161

Multiple Testing and Data Adaptive Regression: An Application to HIV-1 Sequence Data

Merrill D. Birkner, Sandra E. Sinisi, and Mark J. van der Laan

Abstract

Analysis of viral strand sequence data and viral replication capacity could potentially lead to biological insights regarding the replication ability of HIV-1. Determining specific target codons on the viral strand will facilitate the manufacturing of target specific antiretrovirals. Various algorithmic and analysis techniques can be applied to this application. We propose using multiple testing to find codons which have significant univariate associations with replication capacity of the virus. We also propose using a data adaptive multiple regression algorithm to obtain multiple predictions of viral replication capacity based on an entire mutant/non-mutant sequence profile. The data set to which these techniques were applied consists of 317 patients, each with 282 sequenced protease and reverse transcriptase codons. Initially, the multiple testing procedure (Pollard and van der Laan, 2003) was applied to the individual specific viral sequence data. A single-step multiple testing procedure method was used to control the family wise error rate (FWER) at the five percent alpha level. Additional augmentation multiple testing procedures were applied to control the generalized family wise error (gFWER) or the tail probability of the proportion of false positives (TPPFP). Finally, the loss-based, cross-validated Deletion/Substitution/Addition regression algorithm (Sinisi and van der Laan, 2004) was applied to the dataset separately. This algorithm builds candidate estimators in the prediction of a univariate outcome by minimizing an empirical risk, and it uses cross-validation to select fine-tuning parameters such as: size of the regression model, maximum allowed order of interaction of terms in the regression model, and the dimension of the vector of covariates. This algorithm also is used to measure variable importance of the codons. Findings from these multiple analyses are consistent with biological

findings and could possibly lead to further biological knowledge regarding HIV-1 viral data.

1 Introduction

1.1 Motivation

Sequencing a virus, such as the Human Immunodeficiency Virus Type 1 (HIV-1), could potentially give further insight into the genotype-phenotype associations of a virus. The replication ability of a virus is vital, especially in the case of HIV, where replication is proportional to the severity of disease.

A retrovirus such as HIV-1 has a genome consisting of RNA. The virus relies on a reverse transcriptase to perform a type of reverse transcription of its genome from RNA into DNA for insertion and integration into the host's genome, otherwise known as the provirus. The retrovirus itself is a storage place for its RNA and the reverse transcription or viral replication takes place in the cytosol of the virus.

The reverse transcriptase and protease regions of the virus are important areas to consider when studying the viral replication capacity of HIV-1. The reverse transcriptase enzyme works in the capsid and is used to synthesize the double stranded DNA from the virus' single stranded RNA genome. This therefore leads to the viral integration into the host's chromosome, where it becomes the template for RNA virus strands by the host replication machinery, thus, making this region essential for viral replication. The protease is an enzyme that breaks the peptide bonds of proteins. The virus depends on these enzymes in its reproductive cycle, by cleaving nascent polyproteins during viral replication.

Since these two specific regions are important in viral replication, these areas of the virus must be analyzed when researching replication. Many antiretrovirals have been manufactured to target these specific areas of the viral strand. Antiretrovirals, known as protease inhibitors, inhibit the activity of protease, and therefore thwart the process used by the virus to cleave viral proteins. These inhibitors therefore prevent final assembly of the HIV-1 virons. Reverse transcriptase inhibitors are another class of these antiretrovirals. By inhibiting the activity of the reverse transcriptase, they prevent the process of infecting the host's cell. Lack of this enzyme prevents HIV from building pro-viral DNA based on its RNA.

The reverse transcriptase and protease regions of a viral strand must be sequenced when determining specific internal drug target positions. The codon positions of these viral regions are subsequently determined. These codon positions consist of three base pairs, and each individually code for

a specific amino acid. Each codon position could have one of the twenty possible amino acids.

Finally, in order to obtain a correct phenotypic assessment of the virus, a measure of replication capacity is used. Replication capacity is the ability of a virus to replicate in an ideal environment. This is an 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).

Statistical Application to Biological Data

Once sequencing the position specific codons of the viral strand and measuring the replication capacity of the virus, many approaches can be used to determine important codons which may be predictive of the replication capacity of a virus. This data structure lends itself to both a multiple testing procedure as well as an application for a regression based algorithm. Both of these techniques will be applied separately to the data set to determine codons or groups of codons which may predict replication capacity.

Previous analyses on this data set include the application of tree based methods, in particular random forests (Segal et al., 2004). Segal et al. (2004) found tree-structured models to be effective methods when analyzing amino acid based viral sequence data. Their resulting trees illustrated three main codons which were associated with viral replication capacity, which correspond to the first three partitions of their tree. These codons will be discussed in Section 3.

2 Methods

2.1 Data Structure

The HIV-1 sequence dataset consists of 317 records linking the replication capacity (RC) with the reverse transcriptase (RT) and protease (PRO) sequence data from individuals participating in studies at the San Francisco General Hospital and Gladstone Institute of Virology (Segal et al., 2004). The protease positions 4-99 and reverse transcriptase positions 38-223 of the viral strand are used. Each of these codon positions contains three base pairs, which in turn codes for an amino acid. In total there are 282 positions, with a median of 4 amino acids per position.

At each position there are usually a majority of patients exhibiting one amino acid as compared to the other possible amino acids at that position. There are 282 covariate positions with the number of possible amino acids ranging from 1-10. The outcome is a continuous measure of replication capacity ranging from 0.261 to 151.

The positions were coded as binary covariates with each codon position corresponding to either a mutation or no mutation. The value **zero** represents the majority of individuals with one specific codon, and the value **one** represents all of the individuals exhibiting other codons at that position, therefore the minority of individuals. The position $rt178$ was somewhat ambiguous since two of the amino acids were exhibited by a majority of the individuals. In order to accurately code this position, we referred to biological research which indicates that the subtype B consensus amino acid at that position is I (isoleucine) and the mutations are L (leucine), M (methionine), and V (valine) (see http://hivdb.stanford.edu/cgi-bin/RTMut.cgi). This procedure of coding the positions as mutant/non-mutant has been performed in several other analyses of HIV-1 data (Wu et al., 2003; Gonzales et al., 2003). The natural log of the outcome of replication capacity is used in all of the following analyses. This will allow for a more accurate interpretation of the results, since this transformation will decrease the impact caused by the extreme outliers.

We will define the observed data structure for a subject i as $O_i = (Y_i, W_i)$, $i = 1, ..., n$ ($n = 317$ individuals), where $W_i = (W_{1i}, \ldots, W_{pi})$ is a p-vector of explanatory variables (e.g., codon positions), and Y_i is the scalar outcome (e.g., log replication capacity).

2.2 Statistical Analysis

2.2.1 Multiple Testing Procedure

A multiple testing procedure is applied to the data set to test each codon position with the outcome of the natural log of the replication capacity. The testing approach creates a null distribution for the test statistics as opposed to a null data-generating distribution. In the following section we outline the steps used in this procedure, which includes: defining the test statistics, null hypotheses and parameter of interest; defining the error rate which we wish to control; defining the null distribution of the test statistics Q_0 with a resampling/bootstrap approach; and finally creating adjusted p-values for

the M tests.

We perform simple regressions of replication capacity against the specific codon, and estimate $E(Y|W_m)$, where $m = 1, ..., M$ corresponds to the M codon positions. For each regression, the parameter of interest is the coefficient of the codon, β_1. The null hypothesis for each of these M tests is a two-sided hypothesis, since we are interested in large absolute values of the test-statistic. In this case, the null hypothesis is: $H_{0m} : \beta(m) = \beta_0(m)$, with $\beta_0(m) = 0$, and the alternative hypothesis is $H_{1m} : \beta(m) \neq \beta_0(m)$. The corresponding test-statistic is defined as: $T_n(m) \equiv \sqrt{n}\frac{\beta_n(m)-\beta_0(m)}{\sigma_n(m)}$, where $\frac{\sigma_n(m)}{\sqrt{n}}$ is the estimated standard error of $\beta_n(m)$. The hypothesis H_{0m} is rejected if $T_n(m) > c(m)$. $c(m)$ is selected to control a desired Type I error under an appropriate null distribution for $T_n(m)$, where $m = 1, ..., M$. In the process of constructing the test statistics and null hypotheses, we assume that $\beta_n(m)$ is an asymptotically linear estimator of $\beta(m)$ (Pollard and van der Laan, 2003).

First, we define the error rate that we wish to control. We consider three types of Type I error rates: family-wise error rate, generalized family-wise error rate, and tail probability of the proportion of false positives. We are interested in controlling these three error rates, separately, at the 5-percent α-level.

The family-wise error rate (FWER) is defined as the probability of at least one Type I error. The FWER error rate $\theta(F_{V_n})$ is a function of the distribution F_{V_n}, where V_n is the number of false positives, and therefore $FWER = 1 - F_{V_n}(0) = P(V_n > 0)$. The generalized Family-Wise Error Rate (gFWER) is the probability of at least $k + 1$ Type I errors. This error rate is defined as: $gFWER(k) \equiv Pr(V_n > k) = 1 - F_{V_n}(k)$. When $k = 0$, the gFWER is equal to the previously defined family-wise error rate, FWER. Finally, the tail probability of the proportion of false positives (TPPFP) is based on controlling the probability of the proportion false positives (V_n) to the total number of rejected hypotheses (R_n), at a user supplied q and α level. This error rate is therefore a function of the joint distribution of the number of false positives and rejections. $PFP(q) \equiv Pr(V_n/R_n > q)$, where $q \in (0, 1)$.

Once we define the error rates that we are interested in controlling, we generate a null distribution. The estimated null distribution is used to derive a common cut-off value c_o for the test statistics $T_n(m)$ such that a given Type I error rate (described above) is controlled at a specific user defined level α.

Pollard and van der Laan (2003) proposed as the null distribution the asymptotic distribution of the mean-zero centered test statistics, or equivalently the asymptotic distribution of $(\sqrt{(}n)(\beta_n(m)-\beta(m)) : m)$. This null distribution can be estimated with the bootstrap distribution of $T_n^{\#}(m) = \sqrt{n}\frac{\beta_n^{\#}(m)-\beta_n(m)}{\sigma_n^{\#}(m)}$ (where i.e. $\beta_n^{\#}(m)$ corresponds to $\beta_n(m)$ calculated from a bootstrap sample). They proved that with this null distribution the single-step procedures based on the common cut-off rules for each test statistic, $T_n(m)$, provide asymptotic control of any Type I error rate that is a function of the distribution of the number of false positives, V_n (Dudoit et al., 2004). This approach is generalized to general test statistics for general null hypotheses of the form $H_{o,j} : P_o \in M_j$, where M_j is an element of a specific statistical model.

As described in detail in Pollard and van der Laan (2003) and Dudoit et al. (2004), the Bootstrap method can be implemented as follows: Initially, one generates B bootstrap samples, $(X_1^b, ...X_n^b), b = 1, ..., B$. For each bootstrap sample, an M-vector of test statistics is computed, $T_n^{\#}(., b) = (T_n^{\#}(m, b)) : m = 1, .., M$, which is arranged in an $M \times B$ matrix. This matrix will be denoted as $T_n^{\#}$, with rows of this matrix corresponding to the M hypotheses and the columns correspond to the B bootstrap samples. The row means $E[T_n(m)]$ of the matrix T are computed, and the matrix is shifted by the respective mean. The test statistics true distribution $Q_n(P)$ is replaced by a null distribution Q_0, and the bootstrapped estimate of this null distribution (Q_0) is denoted with Q_{0n}. After calculating the bootstrap matrix, one can easily obtain the common cut-offs, c_0, for controlling family-wise error (FWER) control. For FWER control $(k = 0)$, the general procedure is summarized as the single-step maxT procedure, based on the maximum test statistic for each column in the Q_{0n} matrix. The estimated common cut-off value c_o is the $(1 - \alpha)$ quantile of the B-vector of maximum values, obtained from the estimated bootstrapped distribution. This now defines a Multiple Testing Procedure $S_{FWE}(T_n, Q_{0n}, \alpha)$.

Finally, given a Multiple Testing Procedure $S_\theta(T_n, Q_0, \alpha)$, the adjusted p-values for each of the M tests are defined as follows. The adjusted p-value $\tilde{P}_n(m) = \tilde{P}(m, T_n, Q_0)$, for null hypothesis H_{0m}, is defined as:

$$\begin{aligned}\tilde{P}_n(m) &\equiv \inf\{\alpha \in [0,1] : \text{Reject } H_{0m} \text{ at MTP } \theta\text{-level } \alpha, \text{ given } T_n\} \\ &= \inf\{\alpha \in [0,1] : m \in S_\theta(T_n, Q_0, \alpha)\}\end{aligned}$$

We note that the adjusted p-values are defined as $\tilde{P}_n(m) = \tilde{P}(m, T_n, Q_{0n})$,

therefore based on the test statistics and null distribution. Since we are controlling the FWER and using the maxT approach, we can calculate these $\tilde{P}$ values from the distribution of the vector of maximum values.

Once we obtain the adjusted p-values controlling FWER, we used simple augmentation techniques to control generalized family wise error (gFWER) and the tail probability of the proportion of false positives (TPPFP)(van der Laan et al., 2004). Therefore we will have three separate adjusted p-values for each test, corresponding to FWER, gFWER(k), and TPPFP(q) control.

In order to control the generalized family wise error, the simple augmentation procedure consists of initially ordering the M FWER adjusted p-values from smallest to largest, $P_{0n}(O_n(1)) \leq \cdots \leq P_{0n}(O_n(M))$, with $O_n(m)$ denoting the indices for the ordered unadjusted p-values $P_{0n}(m)$. The augmentation procedure used to control $gFWER(k)$ will set the first k ordered FWER adjusted p-values equal to 0 and the ordered FWER adjusted p-values of $m = k+1, ..., M$ will result in the offset (by k) of the ordered FWER adjusted p-values. Therefore the $k+1$ ordered FWER adjusted p-value is then equal to the first ordered FWER adjusted p-value, the $k+2$ ordered FWER adjusted p-value is equal to the second ordered FWER adjusted p-value, and so on until the M^{th} ordered FWER adjusted p-value (equal to the $M-k$ ordered FWER adjusted p-value). This process can be summarized as follows:

$$\tilde{P}^{+}_{0n}(O_n(m)) = \begin{cases} 0, & \text{if } m = 1, ..., k, \\ \tilde{P}_n(O_n(m-k)), & \text{if } m = k+1, ..., M. \end{cases} \tag{1}$$

Additional information on this augmentation can be found in van der Laan et al. (2004).

When controlling the tail proportion of the number of false positives (TPPFP), the ordered FWER adjusted p-values (defined above) are again used with a user defined q or proportion of false positives to total rejections. The m^{th} ordered FWER adjusted p-value is shifted by $m \times q$, instead of $m - k$ as indicated in the gFWER procedure. Again, in more formal terms, the augmentation procedure produces adjusted p-values that are defined as (van der Laan et al., 2004):

$$\tilde{P}^{+}_{n}(O_n(m)) = \tilde{P}_{0n}(O_n(\lceil (1-q)m \rceil)), m = 1, \ldots, M. \tag{2}$$

2.2.2 D/S/A Algorithm

The Deletion/Substitution/Addition (D/S/A) algorithm (Sinisi and van der Laan, 2004) is a data-adaptive regression method to predict the conditional expectation of an outcome or response, Y, given a set of inputs or explanatory variables, W, where W is a d-dimensional vector. The goal of the D/S/A algorithm is to estimate $E(Y|W)$ and is completely defined by the choice of loss function, the choice of basis functions, and the sets of deletion, substitution, and addition moves. When applying the D/S/A algorithm to linear regression, we are using the squared error loss function with tensor products of polynomial basis functions. An example of a basis function in this context is the three-way interaction $W_1W_2W_3$.

When running the algorithm, certain fine-tuning parameters are to be selected via cross-validation. We are selecting the size of the model (k_1), i.e., the number of tensor products, the maximum order of interaction for each tensor product (k_2), and the dimension of the vector of covariates (k_0) using v-fold cross-validation. The algorithm splits the data, or learning set, into a training set and validation set, and it builds estimators for k_0, k_1, and k_2 on the training set and evaluates, i.e., computes a cross-validated risk, for these estimators on the validation set.

To use the D/S/A algorithm for polynomial regression, the user feeds the data (Y, W) and specifies a set of constraints over which to search: $k_0 = \{1, \ldots, K_0\}$, $k_1 = \{1, \ldots, K_1\}$, and $k_2 = \{1, \ldots, K_2\}$. If we want to use all d covariates and not reduce the data, then the user can set k_0 to d, a single value representing the number of available covariates. Otherwise, we can look at a set of possible values for k_0. To reduce the number of covariates, we compute d T-statistics corresponding to the main effects of $W_1, \ldots, W_d$ by fitting d univariate regressions. Next, we rank these statistics, possibly in absolute value, in decreasing order $\hat{R}(1), \ldots, \hat{R}(d) \subset \{1, \ldots, d\}$ yielding our ordered covariates $W_{\hat{R}(1)}, W_{\hat{R}(2)}, \ldots, W_{\hat{R}(d)}$. Then, one can input the set $(W_{\hat{R}(1)}, \ldots, W_{\hat{R}(k_0)})$, of length k_0, as the vector of covariates into the D/S/A algorithm. This constraint was placed in hopes to eliminate much of the noise and possible competition of variables in the regression models. Therefore we will be including those variables with strong marginal associations with the outcome.

Let $k_0 = \{6, \ldots, 10\}$, $k_1 = \{1, \ldots, 5\}$, and $k_2 = \{2, 3\}$ to ease in the description of the algorithm. First, we rank our T-statistics as described above and keep the six covariates corresponding to the six highest T-statistics

as candidate covariates, and we begin with $k_2 = 2$.

The algorithm starts by fitting a linear regression model with the main term, W_1 or W_2 or, $\ldots, W_{k_0}$, that minimizes the squared error loss. Next, it will begin cycling through a set of deletion, substitution, and addition moves (Sinisi and van der Laan, 2004). At each step of the algorithm, the goal is to find the linear combination of polynomial basis functions which best predicts Y by minimizing the squared error loss function. The algorithm gives preference to deletion moves, then substitution moves, then addition moves. The algorithm will try a deletion move first, and if there is no deletion move which improves the current best residual sum of squares (RSS) for the fit of the same dimension, then the algorithm will try a substitution move. If it can make a substitution move, it will go back and try another deletion move. Otherwise, it will try an addition move. If it makes an addition move, it goes back to trying the deletion moves. If no addition moves can be made, then the algorithm stops. Meanwhile, the algorithm is only making moves such that each tensor product is a main term or two-way interaction ($k_2 = 2$), and it is keeping track of the *best* model for sizes one through five (k_1). Then, we repeat this process where we allow three-way interactions ($k_2 = 3$). This is then repeated for seven (k_0) candidate covariates, then for eight candidate covariates, and so forth.

At the end of this process, the algorithm has produced a three-dimensional table of cross-validated risks for $k_0 = 6, \ldots, 10$; $k_1 = 1, \ldots, 5$; and $k_2 = 2, 3$. We choose (k_0, k_1, k_2) that corresponds to the minimal cross-validated risk and call those values $\hat{k}_0, \hat{k}_1, \hat{k}_2$. The algorithm is now run on the learning set with $k_0 = \hat{k}_0, k_2 = \hat{k}_2$, and the *best* model of size $\hat{k}_1$ is reported.

Sinisi and van der Laan (2004) compared the D/S/A algorithm to other popular regression techniques such as forward selection, Logic Regression, and Multivariate Adaptive Regression Splines (MARS) and found the predictive power of the D/S/A algorithm to be competitive with all three methods.

2.2.3 Variable Importance Measures

In addition to reporting an optimal predictive model, the D/S/A algorithm produces an importance measure for each variable. Sinisi and van der Laan (2004) proposed a derivative-based method to estimate importance measures for individual variables based on the idea of counterfactual variables in the causality literature (van der Laan and Robins, 2003). Measures of variable importance can assist in the identification of a subset of codons for replication

capacity.

Let the data be n observations of (Y, W), where Y is the outcome of interest and W is a d-dimensional vector of covariates for which we would like a measure of importance. Let $h(W) = E(Y|W)$. Sinisi and van der Laan (2004) explain that they are getting a sense of the importance of variable W_j for $j = 1, \ldots, d$ by seeing what happens when $W_j = w_j$ for a given W_j. Let $\hat{h}_b$ represent a b-specific fit obtained by the D/S/A algorithm. Given a particular b-specific fit $\hat{h}_b(W)$ of $E(Y|W)$ for $b = 1, \ldots, B$, let $\bar{h}_{jb}(w) = \frac{1}{n}\sum_i \hat{h}_b(W_{1,i}, \ldots, W_{j-1,i}, w, W_{j+1,i}, \ldots, W_{d,i})$. The importance measure which aims to measure the "causal" effect of W_j for *binary* variables can be estimated as: $\hat{\alpha}_b(j) = \mid \bar{h}_{jb}(1) - \bar{h}_{jb}(0) \mid$. Sinisi and van der Laan (2004) provide a measure for continuous and general discrete variables as well.

The final estimate of the importance measure is then a weighted average of $\hat{\alpha}_b(j)$ across many b-specific fits. Various approaches for obtaining b-specific fits $\hat{h}_b$ can be considered. The approach we employed is to use the fits for all models of size k_1 and $\hat{k}_0$, $\hat{k}_2$ made by the D/S/A algorithm to estimate $\hat{\alpha}(j)$.

Let $S_b \in \{1, \ldots, d\}$ identify the subset of variables used in a b-specific fit. Then for a given variable, its importance measure is estimated across fits as:

$$\hat{\alpha}(j) = \frac{\sum_{b=1}^{B} \hat{\alpha}_b(j) I(j \in S_b) \mathrm{wt}_b}{\sum_{b=1}^{B} I(j \in S_b) \mathrm{wt}_b} \tag{3}$$

In equation (3), wt represents a weight for a particular fit. We let the weights equal $\mathrm{RSS}/(n-p)$ where n represents the number of observations and p represents the number of parameters in the fitted model.

3 Results

Table 1 presents a list of codon positions that we refer to in Tables 2 - 4 and throughout this section. In these tables $pr55$, for example, refers to $Kpr55G/R$ where we are implying that the mutant amino acids for this position are G (glycine) or R (arginine) and the wild type amino acid is K (lysine). When looking at our results, refer to Table 1 to determine which amino acids are mutant or wild type.

Table 1: Codon Specific Amino Acids. *Note: Table only includes codons referenced in Section 3.*

L **pr10** I/F/R/V
K **pr14** R/G
Q **pr18** E/H/L/P
L **pr19** A/I/Q/T/V
V **pr32** G/I
E **pr34** G/K/Q
P **pr39** S/Q/K/T
R **pr41** E/I/K
K **pr43** R/T
M **pr46** I/L
I **pr47** L/V
I **pr54** G/M/S/T/V
K **pr55** G/R
L **pr63** A/D/F/H/P/Q/S/T/V
H **pr69** E/G/K/Q/Y
A **pr71** E/I/K/L/M/T/V
G **pr73** C/I/M/S/T
V **pr82** A/G/I/P/S/T
L **pr90** M/N/P
M **rt41** A/E/G/L
D **rt67** G/K/N/P/S
R **rt83** K/N
K **rt102** L/M/N/Q/R
D **rt121** C/H/P/L/Y
I **rt135** L/M/P/R/S/T/V
T **rt139** E/I/K/N/R
E **rt169** D/I/K/L/S
Q **rt174** E/H/K/R
D **rt177** E/G/N/P
I **rt178** L/M/P/V/D
M **rt184** E/Q/V/Y
R **rt211** A/G/K/L/M/P/Q/S/T
F **rt214** G/L/W
T **rt215** C/D/F/G/N/S/Y
I **rt202** A/K/S/T/V

3.1 Summary of Previous Results

Sequencing the reverse transcriptase and protease positions of the HIV-1 viral strand is an important method used to determine target areas for antiretroviral therapy. The reverse transcriptase and protease positions facilitate the replication of HIV-1 virus. Sequencing these regions provides biologists with a greater understanding of the genetic mechanism behind the resistance to antiretroviral drugs. HIV-1 drug resistance is the generation of genetic variation in the virus. It is important not only to look at the individual effect of these mutations on the outcome of replication capacity but also to look at potential interaction effects between mutations.

The reverse transcriptase is a DNA polymerase that uses RNA or DNA as its primer. These positions are responsible for producing the double stranded DNA copy of the single strand of RNA found in the virus. This double stranded DNA copy of the viral information can easily be inserted into the host's DNA, therefore facilitating replication. This area of the virus has often been the focus of medical research, since it is the target of drugs, such as AZT, a popular reverse transcriptase inhibitor. Reverse transcriptase inhibitors inhibit the polymerase reaction, which causes the manufacturing of the double stranded DNA. Several codon position mutations are related to antiretroviral resistance and viral replication capacity (positions $rt184$, $rt215$, $rt41$, $rt210$, $rt116$, $rt65$, $rt67$, and $rt69$) (Shafer et al., 2001). 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. The lamivudine-resistance mutation $M184V$ often causes a low-level resistance to the antiretrovirals didanosine and zalcitabine. $Mrt184V$ also reduces replication capacity by reducing the ability of the reverse transcriptase to process correctly. Additionally, mutations at positions of the reverse transcriptase $rt41$, $rt184$, $rt215$ among others have shown resistance to nucleoside analog inhibitors (Shafer et al., 2001).

The protease is an enzyme that is responsible for the post-translation processing of the Gag and Gag-Pol polyproteins, therefore producing the structural proteins and enzymes of the virus. Several 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., 2001). Mutations at several protease cleavage sites also contribute to drug resistance.

Examples of a protease mutation is position $pr10$, where $Lpr10I/F/V/R$ is associated with resistance to all protease inhibitors when present with another mutation. Position $pr90$ has an impact on the substrate cleft of the virus and $L90M$ causes resistance to saquinavir and nelfinavir (protease inhibitors) when combined with various other mutations (Shafer et al., 1998). Position $Ipr54V/L/T$ also causes resistance to the other protease inhibitors when present with other mutations. Mutation $Gpr48V$ has been shown to cause saquinavir resistance, and mutations at residues $pr54$ and $pr82$ produce resistance to indinavir and ritonavir. Mutations in position $pr30$ and $pr90$ in the protease ($Dpr30N$, $Lpr90M$) could cause drug resistance and also reduce the the HIV-1 viral ability to replicate in vitro. Additionally, mutations within $Vpr82A$, $Ipr84V$, and $Lpr90M$ have been associated with a median change in replication capacity (Barbour et al., 2002; Shafer et al., 2001).

Finally, as previously alluded to, replication capacity is a convenient method used to measure a virus' ability to replicate. This measure is often used when assessing the previously mentioned mutations. The method is described as a modification of the phenotypic drug susceptibility test (Barbour et al., 2002). The patient specific gene sequences of the protease and reverse transcriptase are inserted into a virus which contains the luciferase gene (Barbour et al., 2002). The virus is allowed to replicate in this environment and the luciferase activity is measured and compared to the reference virus, which are reverse transcriptase and protease sequences from a known strain of the HIV-1 virus. The HIV-1 virus has been shown from biological research to have a broad range of replication capacity values. These values are useful measures when assessing various aspects of the virus (Barbour et al., 2002).

Segal et al. (2004) applied various methods to data similar to the type used in this article, with 336 observations involving repeated measurements and 276 positions, and provide results given by a tree-structured method (`rpart`), random forests, and logic regression. Segal et al. (2004) eliminated 6, completely conserved sites, of the 282 positions resulting in 276 positions. For our 316 observations, with no repeated measures on a patient, we used all 282 positions with an updated amino acid profile.

For the tree-structured method (TSM) and Random Forests (RF), Segal et al. (2004) used the amino acid specific positions as covariates. Logic Regression (LR) requires the covariates to be binary. Segal et al. (2004) obtained binary predictors by creating contrast indicators for the amino acids

at each position, resulting in 608 indicators. The optimally pruned tree given by the TSM features one split involving $rt184$. In addition, Segal et al. (2004) give a tree with further splits. The top two splits are $rt184$ and $rt215$ which correspond to primary drug resistance sites known to affect replication capacity. Segal et al. (2004) suggest that the third split on $rt178$ is interesting in terms of novelty. Segal et al. (2004) provide a schematic representation of position importance measures for the random forest with minimal prediction error. $rt184$ and $rt215$ are the most prominent where $rt184$ has an importance measure of close to 8 and $rt215$ has an importance measure close to 1. Finally, the fitted LR model is given by one tree with three leaves: $rt184$, $rt215$, and $rt178$. Segal et al. (2004) discuss that though $rt178$ is in the LR model and shown in their TSM, it is not given high importance by RF, and the TSM achieving minimum cross-validated prediction error is that with just one split ($rt184$).

3.2 Multiple Testing Procedure

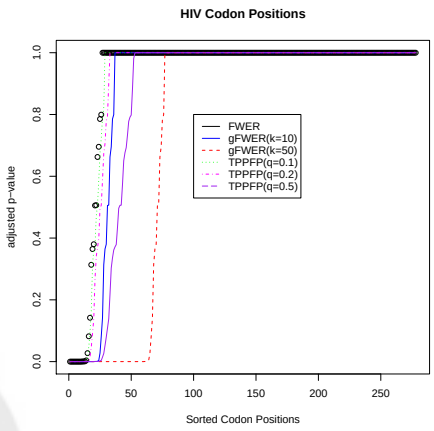

Figure 1: Controlling FWER, gFWER, and TPPFP (sorted adjusted p-values).

Table 2 displays the adjusted p-values controlling for FWER (maxT approach) for 17 codon positions, and Figure 1 plots the sorted adjusted p-values controlling for FWER, gFWER, and TPPFP. The FWER adjusted

Table 2: FWER, PFP, and gFWER controlling adjusted p-values (top 17 codons)

Codon	FWER (max-T)	TPPFP(q=0.1)	gFWER(k=5)
$pr32$	0.0001	0.0001	0
$pr34$	0.0001	0.0001	0
$pr43$	0.0001	0.0001	0
$pr46$	0.0001	0.0001	0
$pr47$	0.0001	0.0001	0
$pr54$	0.0001	0.000108	0.0001
$rt184$	0.00012	0.00012	0.0001
$pr90$	0.00012	0.000127	0.0001
$rt41$	0.00013	0.000131	0.0001
$pr55$	0.00015	0.000149	0.0001
$pr82$	0.0005	0.0004	0.0001
$rt215$	0.0005	0.0005	0.00012
$pr10$	0.004	0.00266	0.00013
$rt121$	0.01	0.008	0.00015
$pr71$	0.028	0.0142	0.0005
$rt102$	0.082	0.0465	0.0005
$rt135$	0.142	0.0967	0.0004

p-values were calculated, as described in the Methods section, and two simple augmentations were separately applied to control the TPPFP and gFWER at a level $q = 0.1$ and $k = 5$, respectively (additional levels of q and k are presented in Figure 1).

The gFWER augmentation, in this case, first orders the 282 FWER adjusted p-values and sets the first five (k) FWER adjusted p-values to 0. The sixth ordered FWER adjusted p-value is then equal to the $m - k$, or first, FWER adjusted p-value, the seventh ordered FWER adjusted p-value is equal to the second FWER adjusted p-value, and so on until the M^{th} ordered FWER adjusted p-value (equal to the 277^{th} FWER adjusted p-value). The TPPFP($q = 0.1$) augmentation procedure first orders the 282 FWER adjusted p-values. The m^{th} ordered FWER adjusted p-value is shifted by $m \times q$, instead of $m - k$ as outlined in the gFWER procedure.

The multiple testing procedure illustrated several codon positions that were significant after controlling for the family wise error rate. Positions such as $rt184$, $rt215$, $rt41$, $pr54$, $pr46$, $pr47$, $pr32$, $pr90$, $pr82$, $pr10$ and $pr71$ have been confirmed in previous work as significant positions with respect to replication capacity and/or antiretroviral resistance (Segal et al., 2004; Shafer et al., 2001). The specific mutations present in our dataset also parallel those found in previous biological research (Table 1). For example, $Ipr54V/L/T$, $Mpr46I$, $Vpr32I$, $Lpr90M$, $Vpr82A/T/F/S$, $Apr71V/T$, and $Lpr10I/F/V/R$ are all protease positions in which mutations increase the resistance to various protease inhibitors. Mutations at several of these positions also have an impact on the replication capacity of the virus. Reverse transcriptase mutation at position $Mrt41L$ increases AZT resistance when present with $Trt215Y/F$. In addition, $Mrt184V/I$ suppresses $Trt215Y$, thus decreasing the AZT resistance (Shafer et al., 2001). This illustrates one of complex mutation processes which occurs between these codons. Other codon positions such as $pr32$ and $pr34$ and $pr54$ and $pr55$ are neighboring codons, respectively, and therefore these mutations, in association with replication, could potentially be of interest for future biological research.

3.3 Results from the D/S/A Algorithm

As a result of the multiple testing procedure, several codons had an adjusted FWER p-value less than the α level of 0.05. This multiple testing procedure created a subset of codons with strong marginal associations with the outcome of replication capacity. We had initially thought of applying the

D/S/A algorithm to this subset of codons, in hopes to eliminate the noise, which is often present in regression models with many predictor variables. We realized that by using this subset, created outside of the D/S/A algorithm's framework, we could be introducing bias in our final model. This is because we are *a priori* predetermining the subset of variables with strong univariate associations (not necessarily choosing these positions because of prior biological knowledge). The bias is introduced since we would not be cross-validating (in the D/S/A algorithm) on the full data, and therefore only a small subset of the data, which was chosen by the multiple testing procedure, thus not performing *honest* cross-validation.

Instead we reduced the number of predictor variables by allowing the D/S/A algorithm to select k_0 with cross-validation, where k_0 is based on univariate associations between a single predictor and the outcome. The D/S/A algorithm was applied to the dataset and cross-validation was used to select: (1) k_1 and k_2 and (2) k_0, k_1, and k_2, as discussed earlier, where k_0 represents the number of initial codon variables to be used as input in the model (dimension of vector of covariates), k_1 represents the size of the final regression model, and k_2 represents the maximum order of interaction for each tensor product. We ran the algorithm such that $v = 5$, in our v-fold cross-validation scheme, $k_0 = 282$ or $k_0 = \{1, \ldots, 282\}$, $k_1 = \{1, \ldots, 10\}$, and $k_2 = \{1, \ldots, 4\}$.

The final results, using the log replication capacity as the outcome, given by the D/S/A algorithm are displayed in Table 3. With constraints imposed only on k_1 and k_2, the final fitted model consists of ten terms ($\hat{k}_1 = 10$) where the maximum order of interaction is two ($\hat{k}_2 = 2$). When we imposed three constraints, the D/S/A algorithm reduced the data to the top nine codons ($\hat{k}_0 = 9$) and produced a model with seven terms ($\hat{k}_1 = 7$) where the maximum order of interaction again is two ($\hat{k}_2 = 2$). This model has potentially important biological implications. First, $rt184$ is illustrated in this model. A $rt184$ mutation is known from previous research (Segal et al., 2004) to be important in the replication capacity of a virus. A mutation in $rt184$ causes the virus to be unable to undergo mutagenesis to reestablish the wild type replication kinetics, and therefore full replication does not occur. The codon $Mrt184V/I$ will decrease replication capacity, as confirmed by the negative coefficient in the models. As mentioned in Section 3.2, many of these codon mutations are biologically important with respect to viral replication capacity and antiretroviral resistance. Protease position $Lpr90M$ is known to have an impact on the substrate cleft and has been shown to

Table 3: D/S/A Algorithm: Fitted Models

$k_0 = 282, \hat{k_1} = 10, \hat{k_2} = 2$
$\log(Y) = 3.649 - 0.715[pr39] - 0.848[rt184] - 3.281[pr10 \times pr32] - 1.737[pr18 \times pr19]$ $+1.601[rt174 \times rt184] - 2.128[pr69 \times rt184] + 0.997[pr63 \times rt184]$ $+0.870[rt184 \times rt202] - 1.772[rt139 \times rt178] + 1.320[rt169 \times rt184]$ $RSS = 148.3$
$\hat{k_0} = 9, \hat{k_1} = 7, \hat{k_2} = 2$
$\log(Y) = 3.648 + 2.063[pr32] - 0.817[rt184] - 0.487[rt41] - 3.206[pr32 \times pr43]$ $+1.038[rt184 \times rt41] - 0.978[pr43 \times rt41] - 2.015[pr47 \times rt184]$ $RSS = 178.7$

confer resistance. The interaction terms which are found in the models could be of biological interest. For example, $Vpr32I$ is a substrate cleft and has been shown to have a minimal effect on resistance. Protease position $pr32$ is interacting with other protease positions in both reported models, which could be of biological interest (Shafer et al., 2001).

The importance measures and proportion of times that codon was used in all the fits made by the D/S/A algorithm for 282 codon positions are displayed in Table 4. Those codons not present in the table had an importance measure of zero. These were calculated for all models fitted by the algorithm for the 282 codon positions using equation (3) for binary covariates. The results (Table 4) highlight $rt139$, $pr32$, and $rt184$ as important variables, among others, within this data set. Again, as mentioned in previous sections, a majority of the mutations illustrated in the table have important biological implications to replication capacity and/or antiretroviral resistance, including a mutation at position $rt184$.

The importance measures for Model 2, or a subset of 9 codon positions are presented in Table 5. This model eliminates some of the noise of the Model 1 regression by including a subset of variables with strong univariate

Table 4: Sorted Variable Importance Measures for Model 1

Codon Position	VIM	Proportion
$rt139$	0.9439	0.31
$pr32$	0.8444	1.00
$pr39$	0.7148	0.08
$rt184$	0.4624	0.85
$pr69$	0.3749	0.54
$rt174$	0.3016	0.62
$pr18$	0.2759	0.77
$rt169$	0.2503	0.15
$pr63$	0.1843	0.46
$rt202$	0.1671	0.38
$pr10$	0.0719	0.92
$pr19$	0.0386	0.69
$rt178$	0.0368	0.23

Table 5: Sorted Variable Importance Measures for Model 2

Codon Position	VIM	Proportion
$pr32$	1.0947	1.00
$rt184$	0.6213	0.89
$pr55$	0.5759	0.05
$pr47$	0.3575	0.42
$pr43$	0.1925	0.95
$pr34$	0.1733	0.16
$pr41$	0.1702	0.79
$pr90$	0.1174	0.58
$pr54$	0	0

associations with the outcome. Again $pr32$ and $rt184$ appear to have high variable importance measures. In both tables, $pr32$ has a proportion of one meaning that it was used in every fit. We eliminate some of the codons which appear in Table 4, which could be noise, and pick up other potentially important codons, such as $pr47$. Mutations in $pr32$ and $pr47$, from previous research, have been shown to cause resistance to the antiretroviral Lopinavir (LPV).

4 Discussion

As illustrated in these analyses, specific codons, or areas on the viral strand, have important univariate associations with the log of the replication capacity or predictive power in determining a virus' replication capacity. We have presented two methods to analyze this viral strand data. The first of which was a multiple testing method, separately controlling FWER, gFWER and TPPFP, which elucidated those codons with strong marginal associations with the outcome of replication capacity. As discussed above, many of these codons with an adjusted p-value less than 0.05, controlling FWER, have been shown to be biologically important with replication capacity and/or antiretroviral resistance. We then applied a data-adaptive regression algorithm, the D/S/A algorithm.

It has been described earlier that Segal et al. (2004) detected $rt178$ with a tree-structure method, coded by its individual amino acids, and with Logic Regression, coded by contrast indicators, but not with Random Forests. With our data, we coded $rt178$ as mutant/non-mutant, based on *a priori* biological knowledge, with the non-mutant amino acid corresponding to the subtype B consensus amino acid. We were interested in the prediction of replication capacity, and we therefore thought that the mutant/non-mutant coding, based on biological information, would be an appropriate coding of our variables. When interpreting results of HIV-1 codon analyses it is always important to keep in mind the biological information of each amino acid, therefore its mutant/non-mutant information based on previous population studies. We did not detect this codon as a significant codon from our multiple testing procedures (FWER, gFWER or TPPFP). We did observe this variable in an interaction term in Model 1 of the D/S/A algorithm. The variable had a low variable importance measure in that model and did not appear in Model 2. Our reasoning behind this is that this variable could merely

be noise in Model 1. It has a higher proportion of mutant/non-mutants as compared to other variables and therefore could have been chosen by the algorithm to predict viral replication capacity, when examining all of the 282 positions. The univariate association of this variable and the outcome is not strong and therefore was eliminated when Model 2 was run on the data.

Our multiple testing procedure did yield an adjusted p-value of less than 0.001 for $rt184$ and $rt215$. The two models displayed from the D/S/A algorithm (Table 3) include $rt184$ and many of the covariates are included in the top 17 codons controlling for FWER (Table 2).

The results presented in this paper will hopefully be of interest to biologists studying the HIV-1 virus. We chose to create binary predictors but could have used discrete predictors by taking into account the specific amino acids at each position. Previous analyses on this type of data involved repeated measurements for several patients. We did not have repeats in our data, however future datasets of this kind may involve repeated measures. The D/S/A algorithm for univariate prediction can be generalized to apply to repeated measure outcomes.

We would like to thank Mark Segal for the use of this dataset as well as providing a detailed description of the data. Merrill Birkner was supported by an NIH Genomics training grant (5 T32 HG00047); Sandra Sinisi was supported by the NIH under grant number R01 GM67233; and Mark van der Laan was supported in part by the NIH under grant number R01 AI055085.

References

Jason D. Barbour, Terri Wrin, Robert M. Grant, Jeffrey N. Martin, Mark R. Segal, Christos J. Petropoulos, and Steven 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.

Sandrine Dudoit, Mark J. van der Laan, and Katherine S. Pollard. Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL `http://www.bepress.com/sagmb/vol3/iss1/art13`. Article 13.

J. H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–141, 1991. Discussion by A. R. Barron and X. Xiao.

Matthew J. Gonzales, Ilana Belitskaya, Kathryn M. Dupnik, Soo-Yon Rhee, and Robert W. Shafer. Protease and Reverse Transcriptase Mutation Patterns in HIV Type 1 Isolates from Heavily Treated Persons: Comparison of Isolates from Northern California with Isolates from Other Regions. *AIDS Reseach and Human Retroviruses*, 19(10):909–915, 2003.

Katherine S. Pollard and Mark J. van der Laan. Resampling-based Multiple Testing: Asymptotic Control of Type I error and Applications to Gene Expression Data. Technical Report 121, Division of Biostatistics, University of California, Berkeley, June 2003. URL `http://www.bepress.com/ucbbiostat/paper121`.

I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic Regression. *Journal of Computational and Graphical Statistics*, 12(3): 475–511, 2003. URL `http://www.biostat.jhsph.edu/~iruczins/publications/publications.html`.

Mark R. Segal, Jason D. Barbour, and Robert M. Grant. Relating HIV-1 Sequence Variation to Replication Capacity via Trees and Forests. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL `http://www.bepress.com/sagmb/vol3/iss1/art2`. Article 2.

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, 2001.

Robert W. Shafer, Mark A. Winters, Sarah Palmer, and Thomas C. Merigan. Multiple Concurrent Reverse Transcriptase and Protease Mutations and Multidrug Resistance of HIV-1 Isolates from Heavily Treated Patients. *Annals of Internal Medicine*, 128(11):906–911, 1998.

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.

Mark J. van der Laan, Sandrine Dudoit, and Katherine S. Pollard. Augmentation Procedures for Control of the Generalized Family-Wise Error Rate and Tail Probabilities for the Proportion of False Positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. URL `http://www.bepress.com/sagmb/vol3/iss1/art15`. Article 15.

Mark J. van der Laan and James M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer, 2003.

Thomas D. Wu, Celia A. Schiffer, Matthew J. Gonzales, Jonathan Tyalor, Rami Kantor, Sunwen Chou, Dennis Israelski, Andrew R. Zolopa, W. Jeffrey Fessel, and Robert W. Shafer. Mutation Patterns and Structural Correlates in Human Immunodeficiency Virus Type 1 Protease following Different Protease Inhibitor Treatment. *Journal of Virology*, 77(8):4836–4847, 2003.