Harvard University

Harvard University Biostatistics Working Paper Series

Year 2003 *Paper* 2

Nonparametric Methods to predict HIV drug susceptibility phenotype from genotype

Greg DiRienzo*

*Harvard University, greg.dirienzo@gmail.com

SIM 1516
pp: 1--14 (col.fig.: Nil)
PROD. TYPE: COM
ED: SNR
PAGN: Vidya – SCAN: Yoga

STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:000–000 (DOI: 10.1002/sim.1516)

Non-parametric methods to predict HIV drug susceptibility phenotype from genotype

A. Gregory DiRienzo[1,*,†], Victor DeGruttola[1], Brendan Larder[2,‡] and Kurt Hertogs[3]

[1] *Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, U.S.A.*
[2] *Virco UK, 184 Cambridge Science Park, Milton Road, Cambridge, CB4 0GA, U.K.*
[3] *Tibotec-Virco NV, Generaal De Wittelaan 11, B-2800 Mechelen, B.E.*

SUMMARY

Medical management of HIV infection requires an understanding of the relationship between viral genetic sequences and viral susceptibility to antiretroviral drugs. Because of the high dimensionality of the data on viral genotype, traditional statistical methods are not well suited for investigating this relationship. We develop non-parametric methods specifically for the setting where high-dimensional data provides a basis for predicting a low-dimensional response variable. Our non-recursive methods proceed in three stages: (i) build models, in a forward-stepwise manner, that predict phenotype response from genotype sequence; (ii) identify specific patterns of amino acid sequence that are most influential in predicting phenotype, and (iii) identify combinations of codons that have either a concordant or a discordant association in the occurrence of a mutation. The methods are applied to a data set provided by the Virco Group that contains protease genome sequences and IC_{50} measurements on a drug from the protease inhibitor class, amprenavir, for 2747 patient samples. From these methods, we were able to identify eight codons from the protease region of the HIV genome that predict resistance to amprenavir, and to determine pairs of codons that tend either to occur together or to preclude the occurrence of the other member of the pair. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: ■; ■; ■; ■; ■; ■

1. INTRODUCTION

1.1. *Background and significance*

Currently, there are 16 drugs approved for treatment of HIV infection in four different drug classes. Viral populations infecting patients under antiretroviral pressure can rapidly develop

[*] Correspondence to: A. Gregory DiRienzo, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, U.S.A.
[†] E-mail: dirienzo@biostat.harvard.edu
[‡] Current affiliation: Visible Genetics Inc.

Contract/grant sponsor: U.S. National Institute of Health; contract/grant number: NIAID 1 RO1 AI51164.

Received November 2001
Accepted January 2003

resistance to drugs in the patients' treatment regimens. In addition, because of the high degree of cross-resistance to drugs within classes, infecting viruses may develop resistance to drugs other than those to which they have been exposed. Because active drugs from at least two classes are required for suppression of HIV infection, it is vitally important to determine the susceptibility to different drugs of infecting viruses before choosing a drug regimen. When only one drug in a patient's regimen is active, the virus is likely to develop resistance to that drug and others of its class, reducing the patient's future treatment options.

Available treatments target regions in the HIV genome that code for reverse transcriptase (RT) and protease enzymes. The genetic code for each enzyme consists of a sequence of amino acids. At each position, or codon, in the amino acid sequence, one of 20 amino acids is present. The RT and protease regions of the HIV genome consist of amino acid sequences of length 494 and 99, respectively. For HIV of the clade or strain most prevalent in the United States and Europe, there is a sequence that virologists believe characterizes the most common genotype of virus infecting treatment-naive patients; this is referred to as the consensus wildtype sequence. When the amino acid present at a codon differs from the corresponding wildtype amino acid, then the sequence is regarded as mutant at that codon. Resistance arises, at least in part, from the development of mutations at different codons in the genotype.

One approach to predicting drug susceptibility makes use of genetic sequences of HIV obtained from plasma; in recent years, the cost of obtaining such sequences has dropped to the level where they can be used in routine clinical patient management. Prediction of susceptibility poses methodologic challenges, however, because of the high dimensionality of the available genetic information. For example, when studying susceptibility to drugs that target the protease enzyme, known as protease inhibitors, the protease sequence is the basis for prediction. With 99 codons in the sequence, there are 99 covariates for prediction, each one being a nominal categorical variable with 20 possible levels.

Some current methods for handling such high-dimensional data are recursive partitioning [1], bump-hunting [2] and neural networks [3]. In particular, recursive partition [4] and neural networks have been shown to be useful techniques for relating HIV genotype to phenotype. A recent exposition of recursive partitioning and bump-hunting methods to peptide-binding data is given in Segal *et al.* [5]. Since the predictor variables we consider are nominal (amino acid), the multivariate adaptive regression spline (MARS) [6] extension of regression trees cannot be used in this setting.

Recently, for the case when comparing two groups (for example, malignant and normal) with respect to the occurrence rate of a mutation at a single codon, Mutter *et al.* [7] have devised a permutational testing scheme to obtain the overall nominal significance level against which the two-sample tests, calculated at each codon, are to be compared. This method thus protects the rate of falsely judging there to be a significant group difference with respect to the occurrence of a mutation at a single codon when making simultaneous inference across all codons.

In this paper, we develop non-parametric methods for investigating the relationship between a high-dimensional predictor variable, genotype sequence, and a low-dimensional response variable, phenotype. In the HIV setting, phenotype can be measured by either an *in vitro* drug susceptibility assay or *in vivo* by the amount of viral RNA in plasma. Although these methods are applicable to either measure, we focus here on the former. The most frequently used *in vitro* drug susceptibility assay is one that estimates the amount of drug required to reduce the replication rate of the virus by 50 per cent, the so-called 50 per cent inhibitory concentration, denoted IC_{50}. The new methods we propose are especially well suited for (i) identifying specific patterns of amino acids that are associated with increased or decreased susceptibility compared to wildtype and (ii) investigating the complex interactions that exist between different codons in the occurrence of a mutation. These methods allow us to explore interactions in the effects of codons in a way that may be more transparent than approaches making use of neural networks.

1.2. Our analysis approach

In order to predict phenotype from genotype, we identify specific genotypic patterns that are associated with increased or decreased drug susceptibility compared to the wildtype virus. The number of possible amino acid values at each codon is not restricted by our method. However, in the data we have analysed, we have found that considering all possible mutations equally can reduce the predictor space at a minimal loss in explanatory power. The model-building process is forward-stepwise and begins by identifying the best model among all those that use as a predictor information from only a single codon; each model is judged based on the proportion of variation of the response (phenotype) explained. Using the same criterion, the next step is to identify the best model among all those models that use as predictors information from two codons. The process proceeds until increases in the number of codons in the model fails to produce a substantial increase in proportion of variation explained. For example, this model building process would stop at K codons when the proportion of variation explained by the best model that contains $K+1$ codons (among all those models containing K codons) is not substantially larger than that corresponding to the best model that contains K codons (among all those models containing K codons). The measure of the proportion of variation of the response variable explained is cross-validated adjusted R^2. The models are saturated in the sense that all possible interactions of codons, or predictor variables, are simultaneously considered.

Our method of analysis is similar in spirit to bump-hunting, in that we seek subsets of the covariate space within which the expected response is significantly different from the expected response from the wildtype sequence. Note that bump-hunting seeks those subsets of the covariate space with expected response significantly different from the overall expected response. Such approaches differ from recursive methods, in which future splitting of the data is conditional on past splits. That is, the way in which the most predictive set of codons at step K of our model building process is chosen in no way depends on the way that the most predictive set of codons at step $K-1$ is chosen.

Associated with the set of codons chosen as the best for prediction is a corresponding set of possible patterns of mutations; each of these patterns has an associated effect in the prediction model. Because the response variables and amino acid sequences are assumed to be independent and identically distributed realizations across subjects, we can show that under the null hypothesis of no association of mutation pattern with response, these effects are statistically independent of one another and asymptotically standard normal. In order to identify patterns of mutations that are associated with increased or decreased drug susceptibility compared to the wildtype pattern, we develop a method for constructing confidence bands for standard normal probability plots; any effect observed to lie outside of the confidence band is significantly associated with response at the desired type I error level.

The interaction among codons in the occurrence of a mutation is of scientific interest, both to aid in drug selection and to understand mechanisms of resistance. Methods have been proposed that investigate the correlation between amino acid position variables [8]. The methods in Bickel *et al.* [8] use a likelihood-ratio test for assessing the independence among codons in the amino acid value and use permutation techniques for inference. We use a different approach to identify groups of codon/amino acid values, among those selected for having a significant association with phenotype, that tend to occur together either more commonly or less commonly than would result from chance. We adapt the techniques described above, for example, the construction a confidence band for a normal probability plot, to identify these groups of codon/amino acid combinations.

Our methods for these types of investigations differ from those of Bickel *et al.* [8] in several ways. The global inference procedures of Bickel *et al.* [8] are *ad hoc* and the associated global type I error level is unknown; it is only known to be bounded for some subsets of codon/amino acid pairs. As a result, these methods are likely to be conservative, that is, not identify significant associations when they exist. Furthermore, their general methodology only handles codon/amino acid pairs, our methods handle any number of codon/amino acid combinations. Bickel *et al.* [8] do develop specific methods to handle sets of three and four codon/amino acid combinations, but these methods require parametric modelling. Finally, whereas their methods ignore the correlation between different codon/amino acid combinations, our inference procedures take into account this correlation structure.

The paper is organized as follows. Section 2 presents notation and describes the model-building process. Section 3 describes how a confidence band for a normal probability plot can be constructed. In Section 4, these methods are adapted to investigate the interdependencies between codons. Section 5 presents an application of these results to a data set from Virco. Finally, Section 6 presents some results from a simulation study.

The MATLAB code for execution of the methods proposed in this paper is available on request from the authors.

2. MODEL AND NOTATION

The value of the amino acid at codon $j=1,\ldots,J$ for individual $i=1,\ldots,N$ is denoted by X_{ij}. Although our methods can accommodate information on the specific mutation (amino acid) that occurs at each codon by transforming the genetic sequences to a vector of indicator variables, for simplicity in the following discussion we consider only whether the amino acid is wildtype or mutant for each codon. We use the notation $X_{ij}=0$ if the amino acid is wildtype at codon j and 1 otherwise. This simplification of the data has been observed to result in a negligible loss in explanatory power in the data that we have analysed, that is, distinguishing between different mutations has not been observed to add much predictability. The scalar response variable, or phenotype, is denoted Y_i, $i=1,\ldots,N$. We assume that the N responses and associated sequences, $(Y_i,\{X_{ij}:j=1,\ldots,J\})$, $i=1,\ldots,N$, are independent and identically distributed.

Our approach conducts analyses in a forward-selection manner. We start with simple models and build them in a stepwise manner as described above. Let J denote the number of codons under consideration; for the protease region, $J=99$. The goal is to find a subset of the J codons containing K codons, $K \ll J$, that capture most of the predictive information available.

For an arbitrary value of K, our algorithm first selects those $M = \binom{J}{K}$ possible combinations of K codons. Each of the M combinations is a possible reduction in the predictor space from J to K.

We consider a fixed value of K and suppress the dependence of notation on K below. There are $L = 2^K$ possible patterns of the sequence $\{X_{ij} : j \in S_m\}$, where S_m is the set of K codons under consideration, $m = 1, \ldots, M$. Each $\{X_{ij} : j \in S_m\}$ equals one of the L possible K-vectors. Let $I_{im\ell}$ equal 1 if the sequence for individual i, $\{X_{ij} : j \in S_m\}$, is equal to the ℓth sequence, $\ell = 1, \ldots, L$, and 0 otherwise. Thus $\sum_\ell I_{im\ell} = 1$. Let $n_{m\ell} = \sum_i I_{im\ell}$ denote the number of individuals in the ℓth cluster for set S_m. The mean value of the response for the ℓth cluster in set S_m is $\bar{Y}_{m\ell} = (1/n_{m\ell}) \sum_{i=1}^{N} I_{im\ell} Y_i$. The predictive model for the mth combination of K codons is

$$Y_i = \bar{Y}_{m1} + \sum_{\ell=1}^{L} I_{im\ell}(\bar{Y}_{m\ell} - \bar{Y}_{m1}) + \varepsilon_{im} \qquad (1)$$

where $E(\varepsilon_{im}) = 0$, $\bar{Y}_{m1} = (1/n_{m1}) \sum_{i=1}^{N} I_{im1} Y_i$ and without loss of generality we take $\ell = 1$ to denote the pattern with K zeros (all wildtype).

The fit of the model can be assessed by cross-validated adjusted R^2, denoted $\bar{R}^2$, where

$$\bar{R}^2_m = 1 - \left\{ \sum_{i=1}^{N} \tilde{\varepsilon}^2_{im}/(N-1-L) \right\} \bigg/ \left\{ \sum_{i=1}^{N} (Y_i - \bar{Y})^2/(N-1) \right\}$$

$m = 1, \ldots, M$, where $\tilde{\varepsilon}_{im}$ is ε_{im} calculated using $\bar{Y}_{m\ell}$ calculated without subject i. An $\bar{R}^2_m$ can be calculated for each of the M sets of K codons and the combination S_{m_0} corresponding to $\max_m\{\bar{R}^2_m\} = \bar{R}^2_{m_0}$ is selected as the best combination of K codons.

Starting with $K = 1$, a value $\bar{R}^2_{Km_0}$ and corresponding set of K codons, S_{Km_0} can be calculated, this is repeated for $K = 2, 3, \ldots$. The value of K for which $\bar{R}^2_{Km_0}$ remains relatively unchanged is selected as an optimal $K = K_0$, with corresponding codon set $S_{K_0 m_0}$. In the next section we propose methods that validly identify sequences whose predicted responses differ from that of the wildtype sequence.

3. CONFIDENCE BAND FOR NORMAL PROBABILITY PLOT

Given an optimal choice of K, K_0, with an optimal choice of codons, $S_{K_0 m_0}$, it is of interest to study the $L = 2^{K_0} - 1$ deviations $\{(\bar{Y}_\ell - \bar{Y}_1) : \ell = 2, \ldots, L\}$, with the set of codons fixed at $S_{K_0 m_0}$. We suppress dependence of notation on $m = 1, \ldots, M$, and now write $\bar{Y}_{m\ell} \equiv \bar{Y}_\ell$, $\ell = 1, \ldots, L$. The distributions of both $\bar{Y}_\ell$ and $\bar{Y}_1$ are approximately normal for n_ℓ and n_1 sufficiently large; in this case, the distribution of the effects $\{(\bar{Y}_\ell - \bar{Y}_1) : \ell = 2, \ldots, L\}$ are also approximately normal. The null hypothesis we test is that the effects, properly standardized, constitute a *sample* of independent observations from a standard normal distribution. In order to test this hypothesis in a valid way, we develop a confidence band for a normal probability plot; any observation that lies outside of this region is deemed to depart significantly from a standard normal distribution at the desired type I error level.

Formally, the hypotheses we test are $H_{0\ell} : E(\bar{Y}_\ell - \bar{Y}_1) = 0$, and the joint null hypothesis $H_0 = \bigcap_\ell H_{0\ell}$, $\ell = 2, \ldots, L$. As a basis for such a test, consider $U_\ell = (\bar{Y}_\ell - \bar{Y}_1)/\hat{\sigma}_\ell$. The $\{U_\ell : \ell = 2, \ldots, L\}$ are asymptotically independent $\mathrm{N}(0, 1)$ random variables under H_0, since

$\hat{\sigma}_\ell^2 = (1/n_\ell + 1/n_1)\hat{\text{var}}(Y_i)$, where vâr denotes sample variance, is a consistent estimate of the variance of $(\bar{Y}_\ell - \bar{Y}_1)$ under H_0, $\ell = 2, \ldots, L$. The independence under H_0 follows because the subsets of individuals in group ℓ_1 and ℓ_2 are disjoint ($\ell_1 \neq \ell_2$). In order to identify departures from H_0 while maintaining the overall type I error rate, we use the following procedure. Let $U_{(\ell)}$ denote the ℓth order statistic and z_ℓ the corresponding percentile from a $\mathrm{N}(0,1)$ distribution, that is, $P\{\mathrm{N}(0,1) < z_\ell\} = (\ell - 0.5)/(L-1)$, where we have used the continuity correction $0.5/(L-1)$. Define the deviation $D_\ell = U_{(\ell)} - z_\ell$. Under the null hypothesis, we obtain through simulation the value $c_L(\alpha)$ that satisfies $P\{\sup |D_\ell| > c_L(\alpha)\} = \alpha$. Thus those D_ℓ with $|D_\ell| > c_L(\alpha)$ are significant departures from H_0 at level α. At each of many, say B, iterations, the simulation procedure generates $L-1$ independent observations from a standard normal distribution and calculates the supremum of the $L-1$ corresponding values of $|D_\ell|$; $c_L(\alpha)$ is taken as the $(1-\alpha)$100th percentile of the B corresponding supremums.

We now provide a brief step-by-step summary of this methodology:

1. Represent amino acid values as indicator variables, for example, 0 if amino acid equals consensus, 1 otherwise. Multiple indictor variables may be used for a given codon to represent the presence of specific mutations.
2. Identify the one codon associated with the greatest cross-validated adjusted R-squared ($\bar{R}^2$) for prediction of response.
3. According to the $\bar{R}^2$ criterion, successively add codons to the model, including in model all possible interactions among codons. This process stops at K codons when the $(K+1)$th codon is not associated with a significant improvement in $\bar{R}^2$.
4. Determine the number, L, of unique genotype patterns arising from these K codons (for example, unique K-vectors of 0's and 1's) in which there are at least five observations.
5. Simulate B independent sets of $L-1$ independent realizations from a $\mathrm{N}(0,1)$. For each of the B sets order the $L-1$ observations and calculate the absolute difference between each ordered realization and z_ℓ, the corresponding percentile from a $\mathrm{N}(0,1)$. For each of the B data sets, find the maximum of these absolute differences and find $c_L(\alpha)$, the $100(1-\alpha)$th percentile of these B maximums.
6. In the data set of interest, order the $L-1$ standardized mean responses, obtaining $U_{(\ell)}$, and find $D_\ell = U_{(\ell)} - z_\ell$, $\ell = 2, \ldots, L$. Patterns significantly associated with an effect on response at the global α significance level are those with $|D_\ell| > c_L(\alpha)$.

4. SUBMODELS: INVESTIGATING INDEPENDENCE

Once an association is made between a set of genotype patterns and response, we want to investigate the relationship among important codon/amino acid pairs. For example, do codon/amino acid pairs occur independently of one another or are their occurrences correlated.

Given a choice of K, there are M possible sets of K codons and for each set there are $L = 2^K$ possible sequences. Fix K and a set of codons, S_m; the dependence of notation on K and m will be suppressed. The probability of the occurrence of the ℓth sequence is denoted by p_ℓ and is consistently estimated by $\hat{p}_\ell = n_\ell/N$. If the K values of $\{X_{ij} : j \in S_m\}$ occur independently of one another, then each of the L sequences $\{x_{\ell j}\}_{j \in S_m}$, where $x_{\ell j} = 0$ or 1, occurs with probability $p_\ell^0 = \prod_{j \in S_m} P(X_{ij} = x_{\ell j})$. We wish to test the joint null hypothesis

$H_0^{(m)} = \bigcap_\ell H_{0\ell}$ where $H_{0\ell}: p_\ell = p_\ell^0$, that the joint probability of each sequence is equal to the product of the marginal probabilities at each codon.

For N sufficiently large, under $H_{0\ell}$, $\hat{p}_\ell\sqrt{N}$ is approximately normal with mean p_ℓ^0 and variance $p_\ell^0(1 - p_\ell^0)$. Thus $U_\ell = \{N^{-1/2}\sum_{i=1}^N (I_{i\ell} - p_\ell^0)\}/[\sqrt{\{p_\ell^0(1 - p_\ell^0)\}}]$ is asymptotically $\mathrm{N}(0,1)$ under $H_{0\ell}$. Define $\hat{U}_\ell$ as U_ℓ except with $\hat{p}_\ell^0 = \prod_{j \in S_m} N^{-1}\sum_{i=1}^N 1(X_{ij} = x_{\ell j})$. Under $H_{0\ell}$, $\hat{U}_\ell$ is also asymptotically $\mathrm{N}(0,1)$ with the $\hat{U}_\ell$ independent across ℓ since the sets of subjects in group ℓ_1 and ℓ_2 are disjoint ($\ell_1 \neq \ell_2$).

Suppose that $K = 2$, that is, we are interested in all two-way associations. For a given pair of codons (j_1, j_2), the $L = 4$ possible combinations of sequences of amino acids are $(1,1)$, $(0,0)$, $(1,0)$ and $(0,1)$. It is easily shown that it is sufficient to consider the one concordant pair $(1,1)$ (or $(0,0)$) and the one discordant pair $(1,0)$ (or $(0,1)$). This is so because if $P(X_{ij_1} = y \mid X_{ij_2} = x) > P(X_{ij_1} = y)$ then $P(X_{ij_1} = 1 - y \mid X_{ij_2} = 1 - x) < P(X_{ij_1} = 1 - y)$, $x = 0, 1$, $y = 0, 1$. Similarly, if $P(X_{ij_1} = y \mid X_{ij_2} = x) = P(X_{ij_1} = y)$ then $P(X_{ij_1} = 1 - y \mid X_{ij_2} = 1 - x) = P(X_{ij_1} = 1 - y)$, $x = 0, 1$, $y = 0, 1$. Thus, to be more efficient, we combine the two concordant groups into one and the two discordant groups into one.

There are $H = M \times L^*$ statistics under scrutiny, L^* from each of the M combinations of K codons, where L^* denotes the number of subsamples after combining concordant and discordant subsamples, $L^* = 2$ for $K = 2$. Let $\hat{U}_{(h)}$ denote the hth ordered statistic and z_h the corresponding percentile from a $\mathrm{N}(0,1)$ distribution. Denote the deviation $\hat{D}_h = \hat{U}_{(h)} - z_h$ and obtain $c_H(\alpha)$ using an analogous procedure to that outlined in the previous section; those $\hat{D}_h$ with $|\hat{D}_h| > c_H(\alpha)$ are significant departures from $H_0 = \bigcap_m H_0^{(m)}$ at level α. Although for a fixed sample of K codons, S_m, the L^* statistics are independent under $H_0^{(m)}$, the test statistics are not necessarily independent across $m = 1, \ldots, M$, as subsamples through m share subjects. This procedure is approximately valid if the set of H statistics are approximately independent. Alternatively, we can estimate the correlation structure of the set of H statistics through the non-parametric bootstrap [9] and compensate for this correlation accordingly.

Suppose that $\hat{\Omega}$ is a consistent estimate of the $H \times H$ correlation matrix of the H test statistics. Let $\hat{U} = (\hat{U}_{(1)}, \ldots, \hat{U}_{(H)})$ and define $\hat{U}^* = \hat{U}\hat{\Omega}^{-\frac{1}{2}}$. Let $\hat{U}^*_{(h)}$ denote the hth ordered statistic. Denote the deviation $\hat{D}^*_h = \hat{U}^*_{(h)} - z_h$. Those $\hat{D}^*_h$ with $|\hat{D}^*_h| > c_H(\alpha)$ are significant departures from H_0 at level α.

Similar testing procedures may be constructed for studying three-way associations conditional on significant two-way associations, and so on.

5. APPLICATION

We apply the methods described above to a data set provided to us from the Virco Group, consisting of sequences of the protease region ($J = 99$) and $\log \mathrm{IC}_{50}$ measurements for amprenavir, a drug of the protease inhibitor class, for $N = 2747$ patient samples. At each codon we code the value of the amino acid as 0 if wildtype and 1 otherwise; discriminating between different mutations does not significantly add to the predictive power of the model.

In our analyses, the response variable Y_i is the ith individual's $\log \mathrm{IC}_{50}$ measurement for amprenavir. The choice for K_0 satisfies the condition that the change in $\bar{R}^2$ from K_0 to $K_0 + 1$ is less than 0.01. For amprenavir, $K_0 = 8$, the corresponding codons are $(32, 46, 54, 71, 82, 84, 88,$

Table I. Significant effects for amprenavir.

X_{32}	X_{46}	X_{54}	X_{71}	X_{82}	X_{84}	X_{88}	X_{90}	D_ℓ	$10^{\bar{Y}_\ell}$	$10^{\bar{Y}_\ell}/10^{\bar{Y}_1}$	n_ℓ
0	1	0	0	0	0	1	0	−1.80	0.42	0.58	15
0	0	0	1	0	0	0	0	−1.64	0.64	0.87	82
0	0	0	1	0	0	1	0	−1.60	0.56	0.77	26
0	0	0	0	1	0	0	1	1.31	1.10	1.51	9
0	0	0	1	1	0	0	1	1.65	1.11	1.51	5
0	0	0	1	1	0	0	0	1.76	1.26	1.72	6
0	1	0	1	1	0	0	0	2.28	1.63	2.23	8
0	1	0	1	1	1	0	1	2.30	7.51	10.25	5
0	1	0	1	1	0	0	1	2.51	1.80	2.46	5
0	1	0	1	0	1	0	0	2.52	11.86	16.20	6
0	1	1	0	1	0	0	1	2.53	4.41	6.03	5
0	1	1	1	1	1	0	0	2.54	7.56	10.32	7
1	1	0	1	1	0	0	1	2.59	7.26	9.91	6
0	0	0	0	0	1	0	1	2.64	3.00	4.10	9
1	1	0	1	1	0	0	0	2.79	3.96	5.41	8
0	1	0	0	0	0	0	0	2.88	1.52	2.08	14
0	0	0	1	0	0	1	1	3.27	1.42	1.94	13
0	0	1	1	0	1	0	1	3.35	4.28	5.85	10
0	1	1	1	0	1	0	1	3.35	6.04	8.25	9
0	1	0	0	1	0	0	0	3.35	4.01	5.48	8
1	1	1	1	1	0	0	1	3.44	6.75	9.22	11
0	0	1	1	0	0	0	1	3.70	2.09	2.85	15
0	1	1	0	1	0	0	0	3.82	2.32	3.17	11
0	0	1	0	1	0	0	0	4.36	1.71	2.34	28
0	0	0	0	0	0	0	1	4.88	1.19	1.62	63
0	1	0	0	0	0	0	1	4.96	2.04	2.79	30
0	0	1	1	1	0	0	0	5.03	1.97	2.69	25
0	1	0	0	0	1	0	1	5.10	5.94	8.12	22
0	1	1	1	1	0	0	0	5.25	3.03	4.13	24
0	1	1	1	1	0	0	1	5.38	3.57	4.87	23
0	0	0	1	0	0	0	1	5.40	1.27	1.73	61
0	1	0	1	0	0	0	1	5.47	1.70	2.32	30
0	0	1	1	1	0	0	1	6.22	1.90	2.60	49
0	1	0	1	0	1	0	1	6.63	4.80	6.56	41
0	0	0	1	0	1	0	1	6.98	3.94	5.37	39

90) and the associated $\bar{R}^2$ is approximately equal to 52 per cent. These codons have previously been observed to be associated with resistance level to amprenavir [10, 11]. We first focus on analysing the $L = 2^8 = 256$ clusters to identify the most predictive sequence patterns. Then we investigate the working relationship between codons.

Out of the 256 possible clusters, 97 are not empty; we restrict attention to the $L = 40$ clusters with cell counts of at least five. The value of the constant $c_{40}(0.05)$ is approximately 1.1. Table I displays those sequences with $|D_\ell| > c_{40}(0.05)$. As shown in Table I, sequences with a mutation at codon 88 are associated with increased sensitivity to amprenavir compared to sequences that are wildtype at all eight codons, even if they also have mutations at 46 (generally associated with resistance) and at 71. Having the wildtype amino acid at 88, N, is associated with increased resistance to amprenavir as compared to the wildtype sequence.

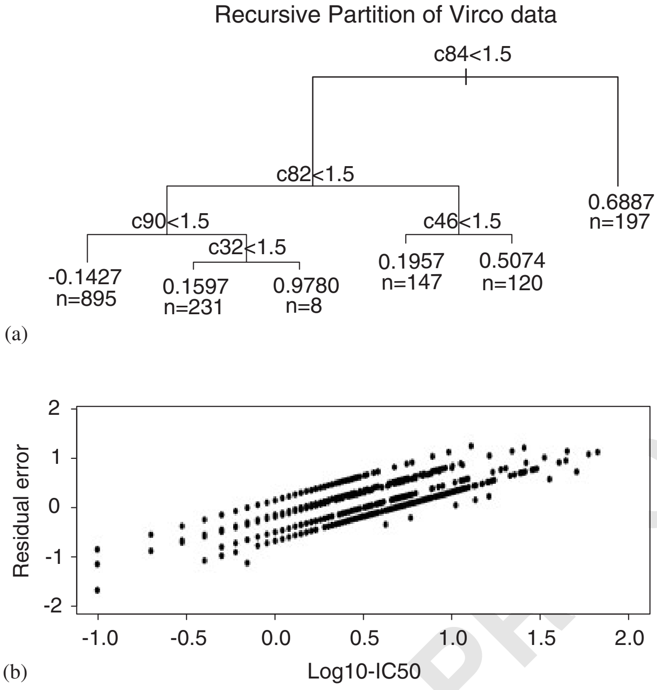

Figure 1. (*a*) Best fit regression tree, wildtype amino acid coded as 1, 2 otherwise, 'c' refers to codon; each terminal node presents the mean $\log_{10} \mathrm{IC}_{50}$ for that node (top number) and node size (n). (*b*) Prediction error of regression tree fit.

Having mutation at two or more of $(46, 54, 82, 84, 90)$ while being wildtype at 88 is strongly associated with increased resistance. Should it be possible to identify a drug that induces a mutation at 88 but preserves wildtype at 84, it might allow sustained sensitivity to amprenavir when used in combination. The (leave-one-out) cross-validated prediction error and $\bar{R}^2$ for this model is 10.7 per cent and 42.3 per cent. The cut-off of 2.5 is used to distinguish between sensitive and resistant virus. Using this 2.5-fold cut-off to distinguish between sensitive and resistant samples, the (leave-one-out) cross-validated estimates of misclassification are 7 per cent of samples misclassified as resistant and 9 per cent misclassified as sensitive.

We also use recursive partitioning methods, with software provided by reference [12], to analyse this data set. The default settings for this software are: a minimal terminal node size of 7; splitting ends when the next split results in a change in R^2 of less than 0.01; and 10 per cent of data is left out for cross-validation. A plot of the resulting pruned cross-validated tree and associated residuals are given in Figure 1. For this model, $R^2 = 46$ per cent, and the adjusted R^2, as defined by this software, is 45 per cent. The corresponding overall misclassification rate is 17 per cent, with 4 per cent misclassified as resistant and 13 per cent misclassified as sensitive. It is interesting to note that all of the codons we identify as being important appear in this tree, but that the codons 88 and 71 did not appear. Because the importance of mutations at 88 and 71 has previously been noted [10, 11], our method appears to provide additional information to that provided by recursive partitioning.

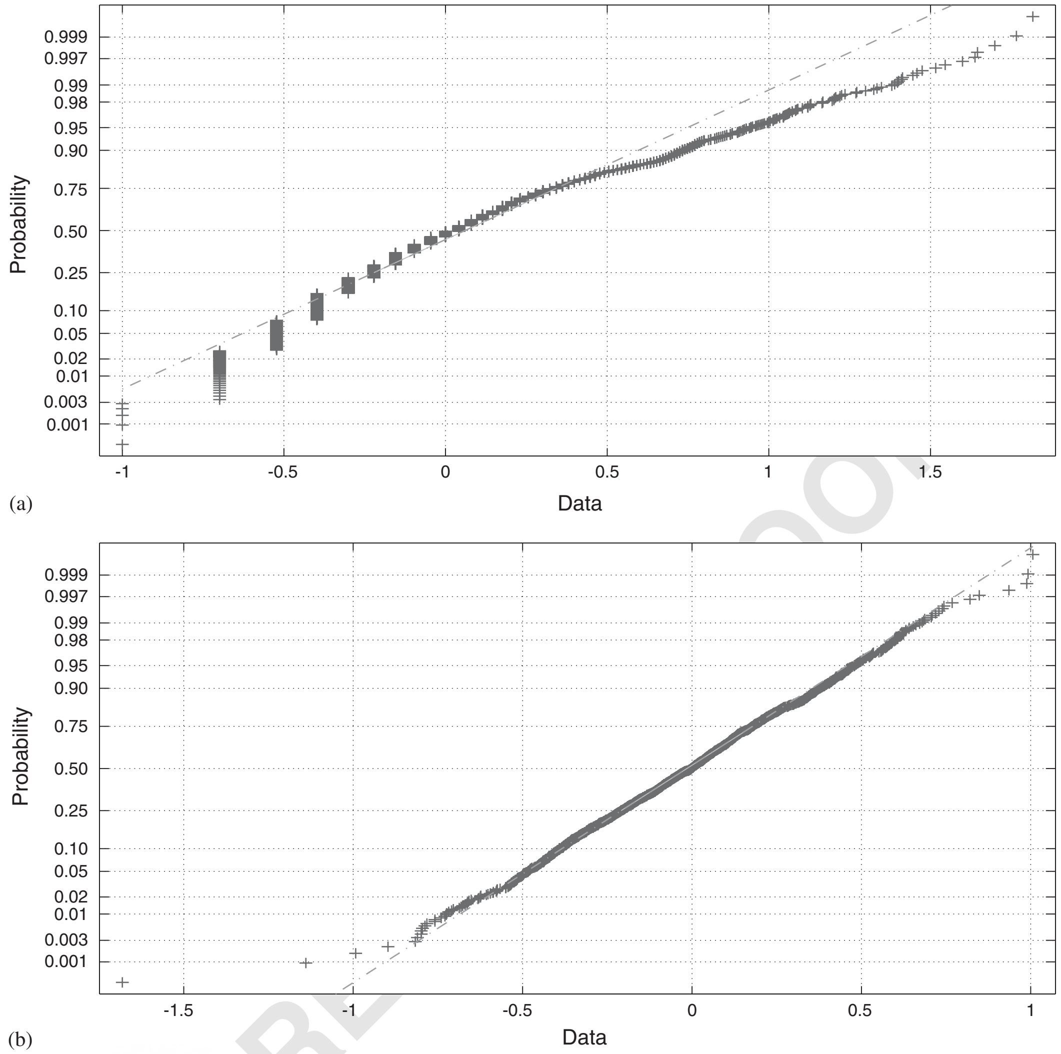

Figure 2. Normal probability plots of (*a*) responses and (*b*) estimated errors.

We next investigate the pairwise association between two codons in the occurrence of a mutation. Significant discordant associations are observed at the following codons: $(88,82)$, $(88,84)$, $(88,90)$ and $(46,54)$. The negative correlation in the occurrence of mutation at codon 88 with mutations at 82, 84 and 90 may be the reason that wildtype at 88 is associated with reduced susceptibility to amprenavir. Significant concordant associations in the occurrence of a mutation are observed at the pairs $(10,84)$, $(32,82)$, $(46,88)$ and $(46,71)$.

Assessment of the assumption of normally distributed errors is based on normal probability plots of the residuals, that is, responses minus respective cluster means (Figure 2(*b*)). Except for a few observations, the residuals show no obvious departure from normality and thus no transformation of the responses is required. The outliers appear to result from some errors in the data, because there are a few observations with many major mutations but no elevation in IC_{50}, and this situation is biologically implausible. A normal probability of the responses is shown in Figure 2(a); given the normality of the residuals, the heavy right tale in this plot appears to result from departures from the null hypothesis.

6. SIMULATION STUDY

We investigate the performance of the proposed methodology in a simulation study. In this simulation, we consider the independent variable, genetic sequence, to be fixed, and simulate responses under several different laws and under both the null and alternative hypotheses.

The choice for K is fixed at $K = 8$ and the codons considered are $(32, 46, 54, 71, 82, 84, 88, 90)$. We only include the 40 genetic clusters with five or more observations, and regard the number of responses in each cluster, or cluster size, as fixed at the number observed in the Virco data. Corresponding to each of these 40 genetic clusters is a cluster-specific mean phenotype, which is used to simulate data under the alternative hypothesis. We fix $K = 8$ because this choice is not a property of our methodology *per se*, but depends rather on the size and complexity of the particular data set at hand. Furthermore, we also consider the set of codons as fixed, choosing the set that maximizes $\bar{R}^2$. Fixing these quantities does not require assumptions about the data generating process. On the other hand, identifying genotype patterns that significantly reduce or increase drug susceptibility relies on the central limit theorem, whose applicability does depend on distribution of responses and on the cell sizes; small cell sizes are particularly a problem when the underlying distribution of response is far from normal.

For the distribution of response, we consider: (i) the uniform distribution on $(-1, 1)$; (ii) the exponential distribution with hazard equal to 1; (iii) the chi-square distribution with 5 degrees of freedom; (iv) the t-distribution with 7 degrees of freedom; and (v) the standard normal distribution. The uniform, exponential, chi-square and t responses are standardized by their respective means and variances, as appropriate, so that all generated data have mean 0 and variance 1. For the null hypothesis, and for each of the 40 genetic clusters, we independently simulate from each of these five distributions a number of independent observations equal to the specific cluster size. The simulated data are used to calculate 39 standardized effects, U_ℓ, $\ell = 2, \ldots, 40$, for each distribution; and the method of Section 3 is used to determine those clusters whose $|D_\ell|$ fall outside of the associated confidence band for the normal probability plot. This simulation was iterated independently 2000 times; the associated asymptotic 95 per cent confidence interval for a true coverage probability of 0.95 is (0.94, 0.96). For each of the five laws for response, Table II provides the percentage of the 2000 iterations in which no effects lay outside the confidence band, which nominally is 95 per cent, as well as the proportion in which one or two or more effects lay outside the confidence band. With the exception of the exponential law, the coverage percentages, for example, the percentages of iterations in which no effects lay outside of the confidence band, are all greater than 90 per cent.

To simulate under the alternative hypothesis, responses are generated from each of the five distributions as above for each genetic cluster; the corresponding observed cluster-specific mean IC_{50} from the Virco data added to each observation. The purpose of this simulation

Table II. Observed percentage of number of effects lying outside of confidence band at 5 per cent type I error level.

Distribution	Null hypothesis			Alternative hypothesis				
	0	1	2+	0	1–9	10–19	20–29	30–39
N(0, 1)	94.5	4.6	0.9	0	0	1.0	31.2	67.8
U(−1, 1)	96.3	3.4	0.3	0	0	0.9	33.9	65.2
exp(1)	83.3	12.0	4.7	0	0	2.1	30.4	67.5
χ_5	90.2	8.0	1.8	0	0	1.3	32.9	65.8
t_7	92.0	7.1	0.9	0	0	0.8	31.7	67.5

is not to study power, but to provide a comparison among these five distributions with respect to their rejection rates under an alternative hypothesis. In Table II, the distribution of the number of effects lying outside of the confidence band is shown for each distribution of response. There is little difference between the five laws in terms of these rejection rates.

These simulations show a fair degree of robustness to distributions that depart grossly from the normal, but in practice, transformation of non-normal response data can improve the performance of the proposed tests. For example, as mentioned above, if observed data are generated by model (1) with independent and identically distributed errors, assessment of normality of the residuals can indicate whether transformations of the responses is necessary. Of course, variance and degree of correlation of residuals within cells depend on the cell size, n_i. Therefore, if cell sizes are small, appropriate transformation of the n_i correlated residuals within a cell into $n-1$ uncorrelated residuals with equal variance across cells may be required before assessing normality.

To examine the benefits of transformations in our simulation, we carry out transformations of response data to produce residuals that are close to being normally distributed. For the simulations based on the exponential law, a Box–Cox power transformation with $\lambda = 1/4$ of the response data (calculated without knowledge that the responses are exponentially distributed) yields nearly normally distributed residuals; the coverage percentages associated with this transformed data are (94.5, 4.8, 0.7). Similarly, for the chi-square law, the associated Box–Cox transformation is $\lambda = 0.3$ with corresponding coverage percentages of (94.4, 4.9, 0.7). From the simulation, it appears that our method is fairly robust to some degree of non-normality of responses, and that appropriate transformations can improve the performance of our method when responses are highly non-normal.

7. DISCUSSION

The proposed methods allow a completely non-parametric investigation of the relationship between patterns of mutations and phenotype; these methods assume no functional form for this relationship. Therefore, for a fixed K, it is not guaranteed that all of the possible clusters of genotype patterns will have enough data points to ensure reasonably accurate estimation of means. As the number of clusters of interest increases, so does complexity, thereby complicating interpretation of results. Because of these properties, our methods are particularly well suited for exploratory analyses, in which avoidance of influential structural assumptions is essential.

The methods developed are appropriate for a high-dimensional setting where the data set of interest has more observations than predictors. Even so, the central limit theorem may not assure distributions sufficiently close to normality when clusters have only a small number of responses. Therefore, it is important to assess normality of the responses, as discussed in Section 6, and to consider appropriate transformations.

Although our method relies on cluster means to be normally distributed, it does have several advantages. While our methods are computationally intensive compared to some tree-based methods, they are much less intensive than neural network approaches. Another advantage is the ease of calculating power to detect cluster effects of different magnitudes for different values of K. These calculations help in interpretation of results, because they provide insight into the amount of information required to support inference of given level of complexity. Such calculations are essential for determining how large patient cohorts must be in order to develop reliable clinical interpretations of HIV genotype. DiRienzo and DeGruttola [13] use the methods developed in this paper to estimate the number of observations needed to detect with adequate power moderate effects of genotype sequence on HIV-1 RNA response.

There exist limits of quantification of the IC_{50} assay, resulting in possibly censored measurements. One way to compensate for this possible censoring is to discretize the $\log_{10} \mathrm{IC}_{50}$ values into the finest partition that eliminates censoring. We could treat such responses as ordered categorical data and make use of a Wilcoxon test for comparison of clusters to wildtype. Analyses would proceed as we described, but using normalized Wilcoxon statistics rather than differences in means. Thus, the flexible approach we describe can be useful in a wide variety of data settings and only requires that the test statistic used for group comparison is asymptotically normal.

ACKNOWLEDGEMENTS

This work was supported in part by grant NIAID 1 R01 AI51164 from the U.S. National Institutes of Health. The authors are grateful to the editor, an associate editor and two referees, whose comments and suggestions led to an improved version of this paper.

REFERENCES

1. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees.* Wadsworth: 1984.
2. Friedman JH, Fisher NI. Bump hunting in high-dimensional data. *Statistics and Computation* 1999; **9**:123–143.
3. Hastie TJ. Neural networks. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: 1998; 2986–2989.
4. Sevin AD, De Gruttola V, Nijhuis M, Schapiro JM, Foulkes AS, Para MF, Boucher CA. Methods for investigation of the relationship between drug-susceptibility phenotype and human immunodeficiency virus type-1 genotype with applications to AIDS clinical trials group 333. *Journal of Infectious Diseases* 2000; **182**(1): 59–67.
5. Segal MR, Cummings MP, Hubbard AE. Relating amino acid sequence to phenotype: analysis of peptide-binding data. *Biometrics* 2001; **57**:632–643.
6. Friedman JH. Multivariate adaptive regression splines. *Annals of Statistics* 1991; **19**:1–67.
7. Mutter GL, Baak JPA, Fitzgerald JT, Gray R,Neuberg D, Kust GA, Gentleman R, Gullans SR, Wei LJ, Wilcox M. Global expression changes of constitutive and hormonally regulated genes during endometrial neoplastic transformation. *Gynecologic Oncology* 2001 (in press).
8. Bickel PJ, Cosman PC, Olshen RA, Spector PC, Rodrigo AG, Mullins JI. Covariability of V3 loop amino acids. *AIDS Research and Human Retroviruses* 1996; **12**:1401–1411.

9. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: 1993.
10. Gong Y, Robinson BS, Rose RE, Deminie C, Spicer TP, Stock D, Colonno RJ, Lin P. In vitro resistance profile of the human immunodeficiency virus type 1 protease inhibitor BMS-232632. *Antimicrobial Agents and Chemotherapy* 2000; **44**:2319–2326.
11. Ziermerman R, Limoli K, Das K, Arnold E, Petropoulos CJ, Parkin NT. A mutation in human immunodeficiency virus type 1 protease, N88S, that causes in vitro hypersensitivity to amprenavir. *Journal of Virology* 2000; **74**:4414–4419.
12. Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the RPART routines. Technical report, Mayo Foundation, Rochester, Minnesota.
13. DiRienzo G, DeGruttola, V. Collaborative HIV resistance-response database initiatives: Sample size for detection of relationships between HIV-1 genotype and HIV-1 RNA response using a nonparametric approach. *Antiviral Therapy* 2002; **7**:593.