Title

Relating HIV-1 Sequence Variation to Replication Capacity via Trees and Forests

Authors

Segal, Mark R
Barbour, Jason D
Grant, Robert M

Publication Date

2004-03-01

Relating HIV-1 Sequence Variation to Replication Capacity via Trees and Forests

Mark R. Segal,[1] Jason D. Barbour,[2] Robert M. Grant[2]

[1] Department of Epidemiology and Biostatistics,
University of California, San Francisco, CA 94143-0560

[2] Gladstone Institute of Virology and Immunology, San Francisco, CA 94141-9100

Abstract

The problem of relating genotype (as represented by amino acid sequence) to phenotypes is distinguished from standard regression problems by the nature of sequence data. Here we investigate an instance of such a problem where the phenotype of interest is HIV-1 replication capacity and contiguous segments of protease and reverse transcriptase sequence constitutes genotype. A variety of data analytic methods have been proposed in this context. Shortcomings of select techniques are contrasted with the advantages afforded by tree-structured methods. However, tree-structured methods, in turn, have been criticized on grounds of only enjoying modest predictive performance. A number of ensemble approaches (bagging, boosting, random forests) have recently emerged, devised to overcome this deficiency. We evaluate random forests as applied in this setting, and detail why prediction gains obtained in other situations are not realized. Other approaches including logic regression, support vector machines and neural networks are also applied. We interpret results in terms of HIV-1 reverse transcriptase structure and function.

Key words: Protease, Random Forests, Reverse Transcriptase, Tree-Structured Methods.

1 Introduction

Genotype-phenotype association studies, wherein genotype is represented by sequence data, are distinguished from standard regression problems by the nature of sequence data. The specific example we investigate pertains to prediction of HIV-1 replication capacity (RC) based on amino acid sequence from reverse transcriptase and protease. A variety of data analytic methods have been proposed in this setting. The methodologic focus is on tree-structured methods and random forests, a recent extension thereof. However, we also briefly critique other approaches that have been used for sequence-phenotype association. The importance of utilizing prediction methods that additionally yield interpretative findings is highlighted by results obtained for RC which admit further insight in terms of HIV-1 reverse transcriptase structure.

1.1 HIV-1 Replication Capacity and Sequence

Replication capacity (RC) of a virus, such as HIV-1, is a measurement of the viruses' ability to replicate in an ideal environment. This would be an environment with abundant cellular targets, no exogenous or endogenous inhibitors, and no immune system responses against the virus. Our measurement of RC employs a single cycle assay, utilizing a portion of the HIV-1 genome (a *gag/pol* segment coding for 2 of 3 enzymes critical to the viral lifecycle). This assay has been shown to correlate to observed in vivo changes in viral fitness associated with loss of drug resistance mutations, and be associated with immunologic outcomes in HIV-1 infected adults. (Deeks et al., 2001; Barbour et al., 2003).

We have recently observed a wide natural variation in RC among viruses with no genotypic or phenotypic markers of RC. Moreover, we have shown that viruses with low RC are associated with increased CD4+ T cell counts. However, the genetic basis for this variation remains unexplained. Identification of sites influencing this wide variation may provide important information about the HIV-1 viral lifecycle, its virulence (capacity to induce disease), variation in clinical outcomes among HIV-1 infected adults, and also provide new drug targets.

Most antiretroviral agents used in HIV-1 treatment target either reverse transcriptase (RT) or protease (PRO). Mutations which confer resistance to these compounds are found within these proteins. Hence, we determined the amino acid sequence of HIV-1 protease (codons 4 - 99) and reverse transcriptase (codons 38 - 223) via population based sequencing (Visible Genetics, Toronto, Canada). Primary resistance mutations are often associated with decreases in RC. Secondary resistance mutations may either increase or decrease RC, but lead to smaller increases in drug resistance. The relationship of secondary mutations to primary mutations in their impact of viral enzyme function, and thus RC, is complex (Barbour et al., 2002).

We obtained 336 records linking RC with RT and PRO sequence data from adults infected with HIV-1. Patients were recruited from an urban, publicly funded clinic which provides care to persons suffering from HIV/AIDS. Of the 336 records, 195 were drawn from a study of acute HIV-1 infection. The remaining 141 were drawn from studies of chronically infected patients in long-term virologic failure of combination anti-retroviral therapy. Of the 336 genotypes obtained, 112 showed genotypic evidence of resistance to at least one class of drug (Protease Inhibitor, non-nucleoside Reverse Transcriptase Inhibitor, or Nucleoside Reverse Transcriptase Inhibitor). The assembled cohort is more than 90% male and 75% Caucasian, with a median age at entry of 39. We note here that a few individuals contributed more than one record and return to this issue in the Discussion.

1.2 Data Analytic Approaches

We have previously described the effectiveness of tree-structured methods (also termed recursive partitioning) for assessing genotype-phenotype association where genotype is represented by amino acid sequence (Segal et al., 2001). In the next section, we provide a brief recapitulation of some of the salient aspects of tree-structured techniques in the present context of HIV-1 replication capacity. We also provide a critique of alternate approaches to genotype-phenotype analyses that have been employed in a similar setting: linear predictor based models, neural networks (Milik et al., 1998; Resch et al., 2001) and prediction-based classification (Foulkes and DeGruttola, 2002).

One of the principal shortcomings of tree-structured methods is their modest prediction performance, attributable to algorithm greediness and constraints which, while enhancing interpretability, reduce flexibility of the fitted functional forms. Random forests (Breiman 2001a,b) are a recently devised methodology, developed to overcome this shortcoming whilst retaining the interpretative strengths of tree-structured approaches. In Section 3, we briefly overview random forests and provide some additional analysis that anticipates their failure to realize predictive gains for the problem at hand. Section 4 offers detailed results of applying both trees and forests to replication capacity with a briefer treatment of some other approaches. Section 5 contains concluding discussion.

2 Approaches to Genotype-Phenotype Association

A wide variety of biomedical problems can be viewed as attempts at relating genotype to phenotype. Here we focus on the case where genotype is represented by amino acid sequence, as exemplified by our protease (PRO) and reverse transcriptase (RT) data, an excerpt of which is provided in Table 1. We use the terms position, site or covariate interchangeably to designate the aligned sequence codons. For continuous phenotypes (outcomes) such as replication capacity, a seemingly natural approach to eliciting relationships would be via some regression procedure. But, as detailed in Segal et al., (2001), and revisited in the present context below, the use of standard (linear predictor based) regression techniques is problematic irrespective of outcome type. Alternate methods, applied to and/or developed for, genotype-phenotype studies also have their limitations as subsequently described.

2.1 Linear Predictor Based Methods

The reason why linear predictor based methods are problematic is the proliferation of indicators needed to encode each covariate (position) and attendant interactions. This arises since (i) the amino acids at any position are inherently unordered with numerous (potentially 20) levels,

(ii) there are numerous sites, and (iii) we anticipate that many phenotypes will be influenced by interactions between sites owing to structural considerations. Huang et al., (1998) provide illustration of this last point in the closely related setting of HIV-1 drug resistance phenotypes and RT structure. Accommodating all possible low-order (second or third) interactions is prohibitive as we illustrate next.

To make these concerns concrete we detail indicator requirements for our dataset of 336 records with paired sequence and RC data. Consensus sequence of protease from positions 4 to 99 and reverse transcriptase from positions 38 to 223 was determined. The resultant total of 282 positions is reduced to 276 following elimination of 6 completely conserved sites. While positions are not generally very polymorphic – the median number of amino acids per position is 3 – we nonetheless require a total of 608 indicators for the simplest model that includes all positions. Since this exceeds the number of individuals, some form of covariate selection and/or regularization (Hastie et al., 2001) will be required. We describe results of applying a regularized method, support vector machines (SVMs), in Section 4.3. Because the use of indicators results in the loss of position integrity otherwise routine tasks, such as appraising individual position and/or amino acid importance, grouping amino acids with similar effects within position, and comparing across positions, become difficult with this many indicators.

But the real limitation comes in attempting to extend to models incorporating interactions due to the explosion in the number of indicators needed. For a set of sequences of length k of an L level residue with each level represented at all positions we require (a) $\binom{k}{2}(L-1)^2$ indicators for all second order interactions; and (b) $\binom{k}{3}(L-1)^3$ indicators for all third order interactions. Here, with variable L (range 2 to 11) and $k = 276$ non-conserved sites, we require (a) 183841 and (b) 36859804 indicators respectively. This precludes fitting using standard statistical software, even when forward selection schemes are employed.

In view of these concerns alternate methods for genotype-phenotype association have been advocated. We briefly critique these next.

2.2 Artificial Neural Networks

In analyzing peptide binding to the class I MHC molecule, K^b, Milik et al., (1998) employ artificial neural networks (ANNs) using bio-physico-chemical properties of the amino acids constituting the peptide. The amino acids themselves were not used for reasons similar to the above indicator proliferation concerns. But which bio-physico-chemical properties to include becomes an issue, with the potential for information loss. Relatedly, in genotype based prediction of HIV-1 co-receptor usage, Resch et al., (2001) use ANNs but (somewhat arbitrarily) score amino acids and gaps on a 1-21 scale. While they contend that results were not sensitive to the scoring system employed, checking this is difficult in view of the $21! \approx 5.11 \times 10^{19}$ possibilities. As we describe below, the virtue of tree and forest techniques is that such scoring schemes are not necessary – unordered covariates (amino acids) are handled with maximal flexibility.

But, there are some more general considerations pertinent to ANNs in the context of genotype-phenotype association. As "black-box" predictors, ANNs are most effective in settings with high signal-to-noise ratios where prediction, not interpretation, is the goal. Such interpretation, in particular identification of important sets of sites is central to genotype-phenotype association studies. Simplistic attempts at "variable importance" extraction using ANN connection weights have been made. However, these typically suffer from profound identifiability concerns arising from the existence of numerous local minima and attendant instability; see Hastie et al., (2001). More refined approaches to interpretation have been recently advanced. Interestingly, in view of the approaches we subsequently examine, these make recourse to tree-structured post-processing (Faraggi et al., 2001) or randomness injection (Intrator and Intrator, 2001). ANN results are presented in Section 4.3.

2.3 Prediction Based Classification

Foulkes and DeGruttola (2002) develop methodology they call prediction based classification (PBC) in the context of predicting phenotypes (drug resistance) from genotypes (HIV-1 sequence). PBC is an elaborate, three-part scheme with the following components. First, individuals are grouped into clusters by applying a clustering algorithm to the sequence data. Second, a recursive partitioning of these clusters into superclusters is performed so as to capture phenotype variation. This is effected using a new construct, the "probability of misordering" (POM) as a split criterion. In turn, the POM is estimated using mixed linear models. Finally, an attempt is made to characterize the resulting superclusters in terms of specific mutations.

Each component of PBC is intricate in its own right. Furthermore, it is seemingly difficult to perform sensitivity analyses since it is difficult to automate or bundle these components. Such sensitivity analyses are important in the context of the PBC approach in view of the many inputs (e.g., distance metrics and clustering algorithm for step 1) for which there are many choices. Here we limit discussion to issues surrounding step 2.

The POM is defined as the probability that an observation from one group of clusters has a higher phenotype value than an observation from another group, conditional on the correct model linking genotype to phenotype. The reason for having to proceed beyond the "correct model" is to allow analysis on the cluster level. The clustering, in turn, was introduced to handle the complexities of sequence data. But, as we describe in Sections 2.4 and 3, there are direct methods for dealing with sequence. The "correct model" is chosen from a hierarchy of linear mixed models of the following form:

$$y_{i(c)} = \beta_0 + b_{0c} + d_{i(c)}(\beta_1 + b_{1c}) + \epsilon_{i(c)}$$

where $y_{i(c)}$ is the resistance phenotype, $d_{i(c)}$ is the distance from wildtype (i.e., number of mutations), and $\epsilon_{i(c)}$ is normally distributed random error for the i^{th} individual in the c^{th} cluster. The unknown parameters (β's and b's) represent fixed and random slopes and intercepts, and are estimated via REML. While considerable effort is invested in exploring competing covariance structures, the mean structure is unchanged over all models examined. This seems like misplaced emphasis given their respective contributions to POM. Furthermore, a mean function that is linear in the number of mutations is inappropriate for modeling resistance phenotypes because (i) primary mutations at active binding sites are more consequential than secondary mutations, (ii) as mentioned above some mutations are compensatory and do not contribute to resistance, and (iii) interactions between sites are not accommodated.

2.4 Tree-Structured Methods

Tree-structured methods (TSM) have seen some recent usage for genotype-phenotype association problems, with even some instances in the HIV-1 context (Beerenwinkel et al., 2002; Quigg et al., 2002). Arguably, the reason for this is the advantages conferred by such methods. However, there are also deficiencies. We discuss both strengths and shortcomings.

The definitive reference describing TSM is Breiman et al., (1984). A review of subsequent extensions to, and refinements of, the basic paradigm is provided by Segal (1995). As mentioned, difficulties encountered by linear predictor based methods occur irrespective of phenotype typology (categoric, continuous, survival). Conversely, TSM deal with all such phenotypes, avoiding these problems by way of the manner in which unordered categorical covariates (e.g., amino acid sequence) are handled. Accordingly, it is this aspect of the methodology that we recapitulate here.

Tree construction involves four components. These are: (1) A set of binary (yes/no) questions, or *splits*, phrased in terms of the covariates that serve to partition the covariate space. A tree structure derives from splitting recursively. The subsamples created by assigning cases according to these splits are termed *nodes*; (2) A *split function* $\phi(s,t)$ that can be evaluated for any split s of any node t which is used to compare competing splits; (3) A means for determining appropriate tree size; and (4) Statistical summaries for the nodes of the tree.

It is the first item that deals with handling covariates. In most implementations (e.g., Therneau and Atkinson, 1997) allowable splits are defined as follows: (a) each split depends upon the value of only a *single* covariate; (b) for ordered (continuous or categorical) covariates, x_j, only order preserving splits of the form "Is $x_j \leq c$?" for $c \in \text{domain}(x_j)$ are considered; (c) for unordered categorical covariates all possible splits into disjoint category subsets are allowed.

So, for the covariate type of interest, unordered categorical, no constraints on possible subdivisions are imposed. If such a covariate has L levels then there are $2^{L-1} - 1$ splits to examine leading to combinatorial explosion for large L. However, by generalizing a result from Fisher (1958), Breiman et al., (1984) establish a theorem that reduces this to an eminently feasible $L - 1$ splits. For the case of continuous phenotypes start by ranking the levels of the unordered categorical covariate by mean phenotype value. Then only those splits that preserve this ranking need be examined. When dealing with the amino acid alphabet, for which in highly polymorphic (variable) settings we have $L = 20$, application of Fisher's result reduces the number of split evaluations per position from a prohibitive 524287 to 19.

It is this *exhaustive* handling of *groups* of amino acids that constitutes one of the advantages of TSM in this setting. Additionally, by the very recursive nature of tree construction, TSM are geared to detecting interactions. A frequently cited deficiency of TSM is that, by virtue of fitting piecewise constant response surfaces, they perform poorly with respect to prediction when faced with smooth response surfaces. This, in part, motivated Friedman's (1991) multivariate adaptive regression spline (MARS) extension of regression trees. However, here such concerns are moot. The very notion of a smooth response surface presupposes the existence of *ordered* covariates – otherwise there is nothing to be smooth with respect to. So, when dealing solely with genotype information represented by unordered categorical covariates this aspect of predictive deficiency does not pertain. We subsequently discuss other aspects of TSM predictive performance.

A final advantage possessed by TSM is the ability to readily provide multiple solutions. This refers to the itemization of both competitor and surrogate splits at each node of the tree; see Breiman et al., (1984) or Segal et al., (2001) for definitions. Loosely, such splits enable the elucidation of alternate, competing models. This is particularly important in the context of genotype-phenotype association since, for a variety of structural and evolutionary reasons, we anticipate strong between-position covariation. A consequence of this dependence is that there will potentially be many competing models having comparable performance. Thus, basing interpretation on a single, optimally chosen model can mask other descriptors. The compilations of competitor and surrogate splits at each node helps offset such overinterpretation. We return to the between-position covariation issue in our evaluation of random forests (Section 3). Results of applying TSM are given in Section 4 and include the identification of a novel site for which there are intriguing structural possibilities whereby RC might be impacted.

Despite these advantages of TSM, pertinent to genotype-phenotype association problems, these techniques have some general deficiencies. The foremost of these is modest prediction performance when compared with more flexible methods, such as ANNs or support vector machines (Cristianini and Shawe-Taylor, 2000). Additionally, TSM are prone to what Breiman terms Rashomon effects (2001a) or instability (1996). These pertain to the widely observed phenomenon whereby small changes in the data and/or algorithm inputs can have dramatic effects on the nature of the solution (variables and splits selected) without affecting predictive performance. Random forests were devised to address these shortcomings.

3 Random Forests

In a series of recent papers, Breiman has demonstrated that consequential gains in classification or prediction accuracy can be achieved by using ensembles of trees, where each tree in the ensemble is grown in accordance with the realization of a random vector. Final predictions are obtained by aggregating (voting) over the ensemble, typically using equal weights. Bagging (Breiman, 1996) represents an early example in which each tree is constructed from a bootstrap (Efron and Tibshirani, 1993) sample drawn with replacement from the training data. The simple mechanism whereby bagging reduces prediction error for unstable predictors, such as trees, is well understood in terms of variance reduction resulting from averaging (Hastie et al., 2001). Such variance gains can be enhanced by reducing the correlation between the quantities being averaged. It is this principle that motivates random forests.

Random forests seek to effect such correlation reduction by a further injection of randomness. Instead of determining the optimal split of a given node of a (constituent) tree by evaluating all allowable splits on all covariates, as is done with single tree methods or bagging, a subset of the covariates drawn at random is employed. Breiman (2001a,b) argues that random forests (i) enjoy exceptional prediction accuracy, and (ii) that this accuracy is attained for a wide range of settings of the single tuning parameter employed.

A random forest is a collection of tree predictors $h(\mathbf{x};\boldsymbol{\theta}_k), \quad k=1,\ldots,K$ where $\mathbf{x}$ represents the observed input (covariate) vector of length p with associated random vector $\mathbf{X}$ and the $\boldsymbol{\theta}_k$ are independent and identically distributed (*iid*) random vectors. We focus on the regression setting (pertinent to RC) for which we have a numerical outcome, Y. The observed (training) data is assumed to be independently drawn from the joint distribution of $(\mathbf{X}, Y)$ and comprises n $(p+1)$-tuples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. The random forest prediction is the unweighted average over the collection: $\bar{h}(\mathbf{x}) = (1/K)\ \sum_{k=1}^{K} h(\mathbf{x};\boldsymbol{\theta}_k)$.

As $k \to \infty$ the Law of Large Numbers ensures

$$E_{\mathbf{X},Y}(Y - \bar{h}(\mathbf{X}))^2 \to E_{\mathbf{X},Y}(Y - E_{\boldsymbol{\theta}} h(\mathbf{X};\boldsymbol{\theta}))^2. \tag{1}$$

The quantity on the right is the prediction (or generalization) error for the random forest, designated PE_f^*. The convergence in (1) implies that random forests do not overfit.

Now define the average prediction error for an individual tree $h(\mathbf{X};\boldsymbol{\theta})$ as

$$PE_t^* = E_{\boldsymbol{\theta}} E_{\mathbf{X},Y}(Y - h(\mathbf{X};\boldsymbol{\theta}))^2. \tag{2}$$

Assume that for all $\boldsymbol{\theta}$ the tree is unbiased, i.e., $EY = E_{\mathbf{X}} h(\mathbf{X};\boldsymbol{\theta})$. Then

$$PE_f^* \leq \bar{\rho} PE_t^* \tag{3}$$

where $\bar{\rho}$ is the weighted correlation between residuals $Y - h(\mathbf{X};\boldsymbol{\theta})$ and $Y - h(\mathbf{X};\boldsymbol{\theta}^{'})$ for independent $\boldsymbol{\theta}, \boldsymbol{\theta}^{'}$.

The inequality (3) pinpoints what is required for accurate random forest regression: (i) low correlation between residuals of differing tree members of the forest, and (ii) low prediction error for the individual trees. Further, the random forest will, in expectation, decrease the individual tree error, PE_t^*, by the factor $\bar{\rho}$. Accordingly, the randomization injected strives for low correlation.

The strategy employed to achieve these ends is as follows:

1. To keep individual error low, grow trees to maximum depth.
2. To keep residual correlation low randomize via
   (a) Grow each tree on a bootstrap sample from the training data.
   (b) Specify $m \ll p$ (the number of covariates). At each node of every tree select m covariates and pick the best split of that node based on these covariates.

However, a key concern from the genotype-phenotype perspective in general, and the RC : HIV-1 sequence data in particular, is that the strategy in 1 controls bias but not variance: such maximal trees may be highly unstable and this instability will be reflected in inflated prediction errors. The reason for this being an acute issue for genotype-phenotype problems is that we anticipate that large numbers of sites will not affect phenotype (i.e., are noise variables). This can result in unpruned, maximally grown, individual trees being so profoundly overfit that (prediction error) recovery via aggregation is partial. Compounding this is (i) limited signal in that very few positions are important, and (ii) the abovementioned between position dependencies. These issues are discussed further in Segal (2003).

As operationalized by the random forest software, available from
`http://www.stat.Berkeley.EDU/users/breiman/rf.html`,
the size of the individual trees constituting the forest is controlled by a tuning parameter, `nthsize`. This specifies the number of cases in a node below which the tree will not split, and so determines maximal tree size. For regression forests the default value is `nthsize` = 5, and this is claimed to give generally good results. For classification forests, the default is `nthsize` = 1, asserted to always give good results. However, the user manual asserts that, in large datasets, larger values can be employed for memory and speed considerations with little loss of accuracy. We investigate impact of varying `nthsize` as well as the primary tuning parameter, m, in Section 4. Further, we introduce and evaluate a further tuning parameter,

anticipated to be helpful in situations where deep trees overfit. This parameter, `nsplit`, governs how many splits per tree are allowable. Note that we could try to achieve such control by adaptively setting `nthsize`, however, this is clearly more awkward. Further, identically sized trees obtained from the two approaches can differ in terms of split covariates and/or cut-points.

Further motivation for constraining the number of allowable splits comes from boosting (Freund and Schapire, 1997). There are some claims that boosting represents an instance of a random forest (Breiman, 2001b) and others that the "resemblance of boosting to such ensemble approaches is at best superficial and that boosting is fundamentally different" (Hastie et al., 2001). If we adopt the perspective of the former with some results of the latter, then there is a basis for investigating limiting the splits per tree. The results in question are given in Hastie et al., (2001, Section 10.11) and show dramatic gains from curtailing allowable tree size.

In addition to excellent prediction performance, random forests possess a number of features. These include measures of covariate importance, distinguishing forests from so-called black-box predictors (e.g., neural nets), and accurate, internal estimates of test set prediction error. Both are illustrated in our analysis of replication capacity.

4 Results: HIV-1 Replication Capacity

4.1 Tree-Structured Methods

The pruned tree structure resulting from fitting RC to all 276 nonconserved PRO and RT positions for all 336 observations is displayed in Figure 1. The given tree withstands cross-validation to the extent that its cross-validated prediction error is less than that of a null tree (i.e, no splitting – simply predicting RC by its mean). However, the optimally pruned tree features just the top split. We discuss tree depth and prediction error further in 4.2.

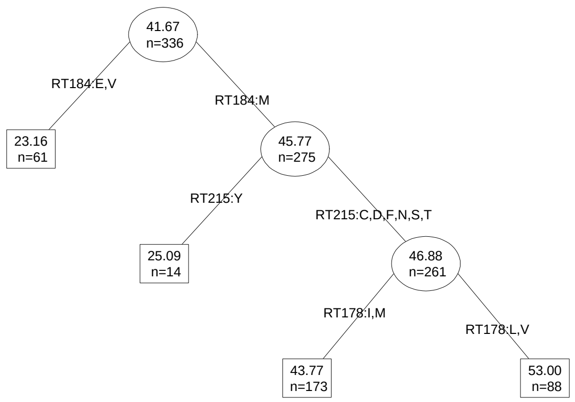

Figure 1: Schematic for (pruned) tree-structured model of replication capacity based on PRO and RT sequence. Within each internal (ellipse) and terminal (rectangle) node the RC average and sample size is given. The branches emanating from internal nodes are labeled with the position and attendant amino acids defining the split.

The top two splits, on RT184 and RT215, correspond to primary drug resistance sites which are known to affect RC. This loosely serves as "proof of principle" re the ability of TSM to extract useful information. Indeed, if we examine the first five competitor and surrogate splits of the root node (i.e., top splits) they all reflect known drug resistance sites for which, again, there are known effects on RC. These splits are detailed in Table 1.

However, it is the third split on RT178 that is the most interesting in terms of novelty. While numerous drug resistance mutations have been shown to lower HIV-1 RC, the role of such naturally occurring polymorphisms in influencing RC has not been explored. RT178 sits along a protein loop that also holds 2 amino acids (D185/D186) critical to the protein's function; see Figure (2). These aspartic acid (D) residues are carefully positioned to coordinate two positively charged metal (Mg^{2+}) ions critical to binding the template. Changes at RT178 away from wild-type may change the position of RT185 and RT186, altering their interaction with the metal ions, and impacting enzyme function.

The RT178 split sorts residues into daughter nodes as follows: Left – Isoleucine (I) or Methio-

Table 1: First splits for the OPTIONS data

Primary Splits		Surrogate Splits	
Split site	Drugs for which site is known resistance mutation	Split site	Drugs for which site is known resistance mutation
RT184	ddI, ddC, ABC, 3TC	PR71	IDV, RTV, SQV LPV
RT215	ZDV, d4T, ABC	PR82	IDV, RTV, possibly LPV
PR90	SQV, IDV	PR90	SQV, NFV
RT67	ZDV, d4T, ABC	RT215	ZDV, d4T, ABC
RT41	ZDV, d4T, ABC	PR73	IDV, SQV, AMP, LPV

nine (M) (mean RC = 43.77); and Right – Leucine (L) or Valine (V) (mean RC = 53). The left daughter node is dominated by I (165 out of 173), while the right daughter node is dominated by V (78 out of 88). As pointed out by a referee, both I and V are hydrophobic, and hence serve similar functions with regard the hydrophobic effect which creates the core structure of folded proteins. However, I occupies a greater volume in the hemisphere defined by rotation around its beta carbon. In addition, one I carbon branch off the beta carbon is one carbon longer than V, and hence provides a greater number of hydrogen bonding opportunities, as well as a different network of hydrogen bonding positions. These volume and hydrogen bonding changes may force a chain of structural changes along the loop containing RT185 and RT186 (which are respectively 7 and 8 residues away from RT178), with similar, albeit attenuated, effect on replication capacity as has been described for the drug resistance substitution M184V that here appears as the first split in Figure 1. Further, while RT185 and RT186 are primary in the interaction with magnesium cations, there are several other sites which participate in coordinating these cations, which may also be disrupted by substitutions at RT178.

4.2 Random Forests

In view of the abovementioned (Section 2.4) modest predictive ability of tree methods, and the potential improvements afforded by random forests (3), we applied the latter to the RC - genotype data. Results with regard prediction error are presented in Table 2. The entries are 10 fold cross-validated estimates of prediction error variance. These closely correspond with

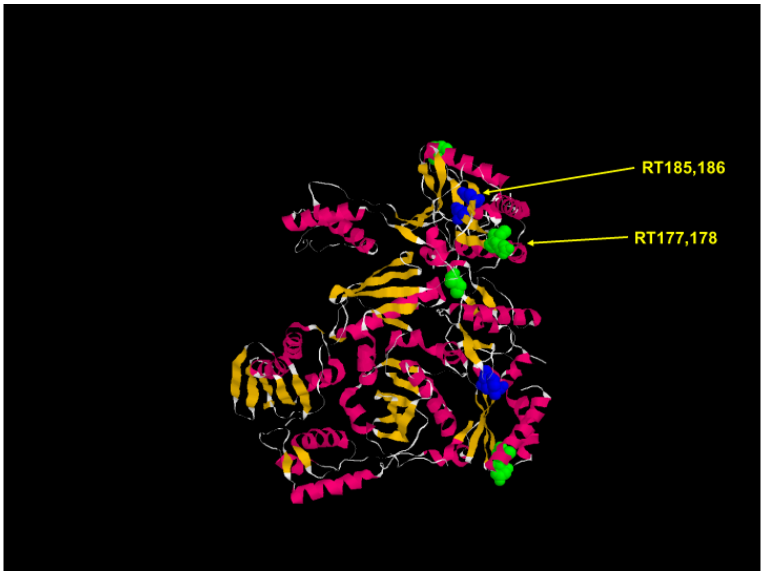

Figure 2: Ribbon diagram of reverse transcriptase identifying key positions.

Table 2: Prediction Errors: Replication Capacity, $n = 336, p = 276$

# Splits per Tree	Minimum Node Size	# Covariates per Split (m) 10	20	100	276
Unrestricted	5	589.7	590.4	608.2	602.9
	25	589.2	586.7	587.5	593.8
	50	594.0	583.7	582.1	584.2
5	5	602.9	592.9	575.6	578.6
	25	598.5	587.4	576.2	577.1
	50	592.4	588.4	581.2	581.6

estimates obtained using 3 or 5 folds and with "out-of-bag" estimates, an accurate estimate of test set prediction error that is "carried along" in ensemble methods employing resampling (bagging, random forests); see Breiman (2001b). Position importances corresponding to the forest with minimal prediction error are depicted in Figure 3. Note consistency with the TSM results in terms of the prominence of RT184 and RT215. However, it is also worth noting that the analog of this measure of variable importance has been superceded in classification random forest software due to volatility when there are many predictors.

Few trends are apparent. For both the recommended default value of m ($\approx p/3 \approx 100$) and for bagging ($m = p = 276$), substantial improvement in PE can be achieved by restricting the number of splits (`nsplit = 5`). For those forests grown without such restriction, gains are realized by increasing the minimum node size (`nthsize`) for which splitting is allowed. Controlling `nthsize` proves unnecessary, or even counterproductive, for the split restricted forests. This illustrates the previously mentioned interplay between `nsplit` and `nthsize` and arguably shows that the former provides a more expedient means for exploring a range of models.

It is of interest to note that the best PE achieved by the suite of random forests examined coincides with the PE attained from a single pruned tree. Such an optimally pruned tree features just a single split. One contributing factor relates to the nature of sequence data. For a variety of biological reasons we anticipate strong between position dependencies. Indeed, applying the likelihood ratio / permutation testing approach developed by Bickel et al., (1996) for assessing correlation when dealing with categorical (e.g. amino acid levels) covariates reveals that, for reverse transcriptase positions, approximately 40% of all possible pairwise position correlations are *simultaneously* significant ($p < 0.01$). It is this strong between position correlation that thwarts the effectiveness of the random forest variance reduction strategy: $\bar{\rho}$ in (3) will be large.

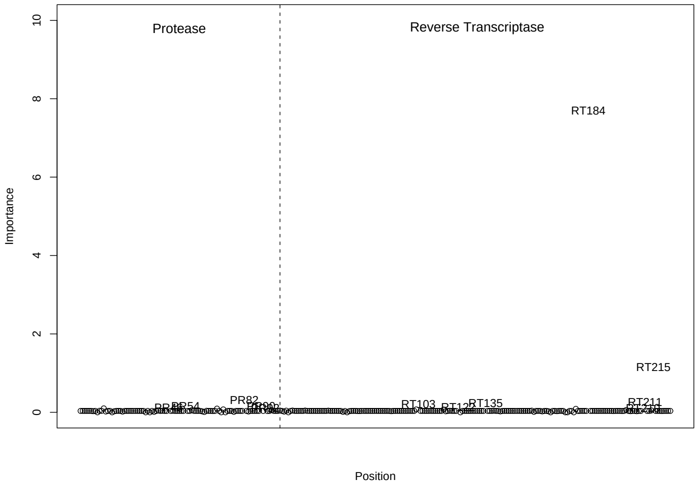

Figure 3: Position importances for the random forest with minimal prediction error.

4.3 Other Approaches

We briefly describe experiences with the application of some alternate prediction tools.

Support Vector Machines

As discussed in Section 2.1, there are fundamental problems in attempting to fit linear models with low order interactions for the RC - sequence data. These derive from indicator proliferation when encoding numerous polymorphic positions and related interpretational concerns. A possible remedy is regularization. We investigated the performance of support vector machines which includes ridge regression (L_2 penalized multiple linear regression) as a special case. The SVM generalizations include basis expansion via kernels and ϵ-insensitive loss functions; see Vapnik (1998) and Cristianini and Shawe-Taylor (2000) for details. We ran dozens of SVM models, focusing on polynomial kernels of degrees 1 through 3, and optimizing other tuning parameters. The best cross-validated prediction error variance (*cf* Table 2) obtained was 595.2. While this investigation was not exhaustive, the failure to improve on the prediction errors of most of the random forest runs or the (single) tree-structured models, coupled with lack of interpretative ease, leads us away from SVMs for this application.

Logic Regression

At the suggestion of a referee, in order to further gauge how “real” the RT178 split is we undertook a logic regression analysis (Ruczinski et al., 2003). Logic regression fits predictive models featuring interactions rendered as boolean combinations of binary predictors. The boolean combinations can be depicted via a tree schematic, enhancing interpretation. The binary predictors were obtained here by creating contrast indicators for the amino acids at each position. Simulated annealing is employed to explore the vast model space. Thus, logic regression has much promise for the problem at hand in view of its flexibile accommodation of interactions and interpretability.

Predictive performance (based on 10 fold cross-validation) for one and two tree models each having up to 7 leaves is shown in Figure 4. The scale used here (y-axis) is prediction standard

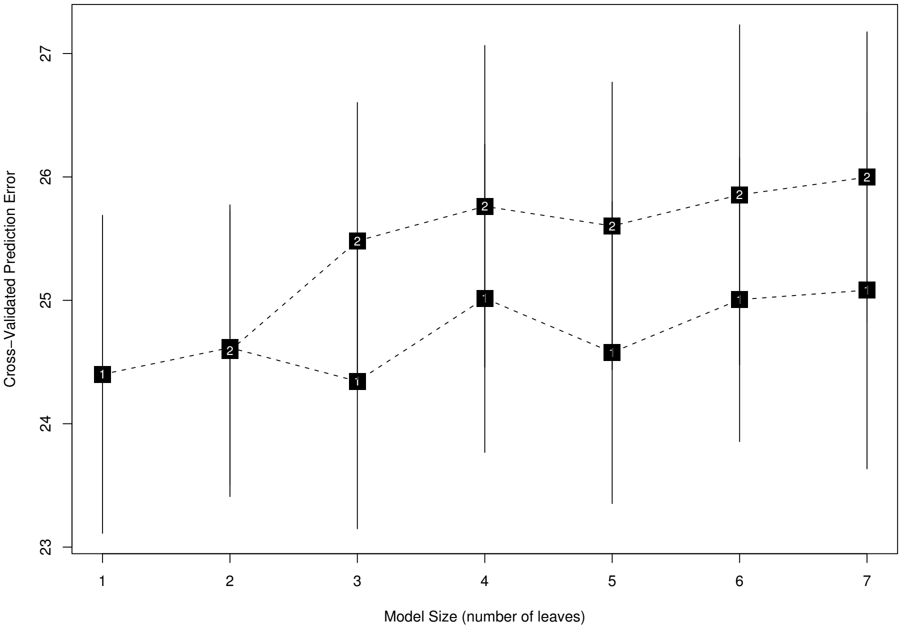

Figure 4: Cross-validated prediction errors for logic regression for RC. Numerals within squares indicate whether a one or two tree logic model was employed. Error bars correspond to ±1 standard error.

error and so it is squared values that are directly comparable to the prediction error variances presented in Table 2. Thus, the optimal logic regression model (one tree with three leaves) has a prediction error variance of $24.34^2 = 592.44$. This exceeds the median prediction error of the differing random forests evaluated per Table 2. Once more, cross-validation exhibits instability, the standard errors bars being large. Indicators representing select contrasts for positions RT184 and RT215 appear most frequently among the best models, the same positions that emerged from the tree and forest analyses. However, RT178 also appears frequently, more so than any other reverse transcriptase position. This selection of RT178 occurs against a pool of 608 indicators and in the context of the vast model space searches explored in fitting logic regression models, and so provides another indication that this position warrants further consideration.

Artificial Neural Networks

We also undertook an investigation of artificial neural networks (ANNs), using the same indicator encoding of sequence positions as for the above methods. While again dozens of models were fitted, we restricted to a single-hidden-layer architecture but allowing skip-layer connections. Regularization was effected using weight decay and this tuning parameter, as well as the number of nodes and other parameters, were varied. Using ANNs we did obtain several instances of smaller cross-validated prediction errors than achieved with random forests. However, the variation of these prediction errors was extreme, dwarfing that for all other methods. Further, following the suggestion of a referee, we tried to measure position (variable) importance by fitting ANNs both with and without position(s) of interest and then differencing the resultant prediction errors. This mimics the variable importance approach used bu random forests. However, this frequently produced negative importances even when

5 Discussion

We briefly summarize methodologic and substantive findings as well as discussing some pertinent surrounding concerns. Firstly, tree-structured methods provide an effective tool for analyzing genotype-phenotype association especially when genotype is represented by amino acid sequence. TSM overcome the fitting and/or interpretational problems affecting other existing approaches. As an aside, it is interesting to note that in order to overcome the interpretational deficiencies of artificial neural networks, both post-processing via tree-structured methods (Faraggi et al., 2001) and randomness injection (Intrator and Intrator, 2001) have been proposed. Secondly, methods intended to improve the predictive performance of TSM, such as forests or bagging, may not yield such gains in this (genotype-phenotype) setting. Reasons for this include the fact that such problems are characterized by limited signal. Few positions (covariates) are important with many being highly conserved; i.e., exhibiting little variation. Relatedly, there are observed and anticipated strong between-position dependencies that serve to partly undermine the random forest strategy. Additionally, sample sizes are typically modest, as was the case for the replication capacity example. It is notable that the benchmark datasets employed in establishing the excellent predictive performance of random forests are (almost) all difficult to overfit using a maximally grown TSM. This sharply contrasts with the present, limited signal setting as is further discussed in Segal (2003).

The use of TSM revealed a putative role for two novel sites as determinants of replication capacity. Polymorphism at RT178 was associated with increased RC. This site lies at the base of the loop containing RT185 and RT186, two of the sites critical for the coordination of Mg^{2+} ions, essential for reverse transcription. An additional TSM analysis, performed on the subset of patients with no genotypic evidence of resistance (as determined by mutations at known resistance sites) and being in early stages of HIV-1 infection (as operationalized by a less sensitive immunosorbent assay that is a proxy for the maturity of an HIV-1 infection; see Janssen et al., 1998) featured a first split at another site, RT135. This was the only split to withstand cross-validation. Mutations at RT135 (i) have been previously implicated as influencing NNRTI resistance (Brown et al., 2000), (ii) may restrict rotation of the RT 'thumb', and (iii) may be influenced by the HLA Class I B5 allele. These analyses and interpretations are further developed and detailed in Barbour et al., (2003). Note the hint of importance ascribed by the random forest analysis (Figure 3).

While we have provided some detailed biologic interpretation surrounding the role of polymorphism at RT178, it is important to recognize limitations surrounding the identification of the corresponding split. In particular, RT178 is not accorded high importance by the random forest and the TSM achieving minimum cross-validated prediction error is that with just one split (RT184). But, the cross-validated prediction error for the tree in Figure 1 with the RT178 split is close, and cross-validation is notoriously unstable (see, for example, Breiman et al., 1984; and Hastie et al., 2001) with the minimum being poorly determined. Additionally, the logic regression analyses suggest that RT178 warrants further consideration.

An open problem from a methodologic perspective is the handling of repeated or serial genotype-phenotype data. In the present data set this facet was not critical since the number of patients furnishing serial data was modest ($<$ 10% of which 83% had only two repeats). Furthermore, the serial genotypes displayed very little variation. Results based on sampling a unique record from those patients contributing multiple observations are concordant with those based on utilizing all 336 records (including repeats) as presented throughout. However, a clearly more satisfactory approach would properly account for the serial data structure. While such methods have been proposed strictly for genotypic changes (Foulkes and DeGruttola, personal communication), the development of techniques that simultaneously handle genotypic and phenotypic change and attendant association is the subject of future research.

Acknowledgements

We thank Virologic Incorporated for providing replication capacity assay results and the referees for several constructive suggestions. This work was supported by NIH grant AI55085. Mr. Barbour is a Burroughs Wellcome Fund Fellow.

References

1. Barbour JD, Wrin T, Grant RM, Martin JN, Segal MR, Petropoulos CJ, Deeks SG. (2002). Evolution of phenotypic drug susceptibility and viral replication capacity during long-term virologic failure of protease inhibitor therapy in HIV-infected adults. *Journal of Virology* **76**:11104-11112.

2. Barbour JD, Hecht FM, Wrin T, Segal MR, Ramstead CA, Liegler TJ, Busch MP, Petropoulos CJ, Hellmann NS, Grant RM. (2003). Higher CD4+ T cell counts associated with low viral *pro/pol* replication capacity among treatment naïve adults in early HIV-1 infection. *Submitted.*

3. Beerenwinkel N, Schmidt B, Walter H, Kaiser R, Lengauer T, Hoffmann D, Korn K, Selbig J. (2002). Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. *Proceedings of the National Academy of Science*, **99**:8271-8276.

4. Bickel PJ, Cosman PC, Olshen RA, Spector PC, Rodrigo AG, Mullins JI. (1996) Covariability of V3 loop amino acids. *AIDS Research and Human Retroviruses*, **12**:1401-1411.

5. Breiman L, Friedman JH, Olshen RA, Stone CJ. (1984). *Classification and Regression Trees*. Belmont: Wadsworth.

6. Breiman L. (1996). Bagging predictors. *Machine Learning*, **24**:123-140.

7. Breiman L. (2001a). Statistical modeling: the two cultures. *Statistical Science*, **16**: 199-215.

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

9. Brown AJ, Precious HM, Whitcomb JM, Wong JK, Quigg M, Huang W, Daar ES, D'Aquila RT, Keiser PH, Connick E, Hellmann NS, Petropoulos CJ, Richman DD, Little SJ. (2000). Reduced susceptibility of human immunodeficiency virus type 1 (HIV-1) from patients with primary HIV infection to nonnucleoside reverse transcriptase inhibitors is associated with variation at novel amino acid sites. *Journal of Virology*, **74**:10269-10273.

10. Cristianini N, Shawe-Taylor J. (2000). *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge.

11. Deeks SG, Wrin T, Liegler T, Hoh R, Hayden M, Barbour JD, Hellmann NS, Petropoulos CJ, McCune JM, Hellerstein MK, Grant RM. (2001). Virologic and immunologic consequences of discontinuing combination antiretroviral-drug therapy in HIV-infected patients with detectable viremia. *New England Journal of Medicine*, **344**:472-480.

12. Faraggi D, LeBlanc M, Crowley J. (2001). Understanding neural networks using regression trees: An application to multiple myeloma survival data. *Statistics in Medicine*, **20**:2965-2976.

13. Fisher WD. (1958). On grouping for maximum heterogeneity. *Journal of the American Statistical Association*, **53**:789-798.

14. Foulkes AS, de Gruttola V. (2002). Characterizing the relationship between HIV-1 genotype and phenotype: Prediction-based classification. *Biometrics*, **58**:145-156.

15. Friedman JH. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, **19**:1-67.

16. Hastie TJ, Tibshirani RJ, Friedman JH. (2001). *The Elements of Statistical Learning.* New York: Springer.

17. Huang H, Chopra R, Verdine GL, Harrison SC. (1998). Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: Implications for drug resistance. *Science*, **282**:1669-1675.

18. Intrator O, Intrator N. (2001). Interpreting neural-network results: A simulation study. *Computational Statistics and Data Analysis*, **37**:373-393.

19. Milik M, Sauer D, Brunmark AP, Yuan L, Vitiello A, Jackson R, Peterson PA, Skolnick J, Glass CA. (1998). Application of an artificial neural network to predict specific class I MHC binding peptide sequences. *Nature Biotechnology*, **16**:753-756.

20. Quigg M, Frost SDW, McDonaugh S, Burns SM, Clutterbuck D, McMillan A, Leen CS, Brown AJ. (2002). Association of antiretroviral resistance genotypes with response to therapy. *Antiviral Therapy*, **7**:151-157.

21. Resch W, Hoffman N, Swanstrom R. (2001). Improved success of phenotype prediction of HIV-1 from envelope variable loop 3 sequence using neural networks. *Virology*, **288**:51-62.

22. Ruczinski I, Kooperberg C, LeBlanc M. (2003). Logic regression. *Journal of Computational and Graphical Statistics*, **12**:475-511.

23. Segal MR. (1995). Extending the elements of tree-structured regression. *Statistical Methods in Medical Research* **4**:219–236.

24. Segal MR, Cummings MP, Hubbard AE. (2001). Relating amino acid sequence to phenotype: Analysis of peptide binding data. *Biometrics*, **57**:632-643.

25. Segal MR. (2003). Machine learning benchmarks and random forests. *Submitted.*

26. Therneau TM, Atkinson EJ. (1997). An introduction to recursive partitioning using the RPART routines. *technical report: Mayo Foundation*

27. Vapnik V. (1998). *Statistical Learning Theory.* Wiley, New York.