Prediction of Human Immunodeficiency Virus Type 1 Drug Resistance:

Representation of Target Sequence Mutational Patterns via an n-Grams Approach

Majid Masso
Laboratory for Structural Bioinformatics, School of Systems Biology
George Mason University
Manassas, Virginia, USA
mmasso@gmu.edu

***Abstract*—Antiretroviral medications for treating human immunodeficiency virus type 1 (HIV-1) infection, in particular inhibitors of the HIV-1 protease (PR) and reverse transcriptase (RT) enzymes, are vulnerable to the emergence of target mutations leading to drug resistance. Here we explore the relationship between PR and RT mutational patterns and corresponding changes in susceptibility to each of their eight and 11 inhibitors, respectively, by developing drug-specific predictive models of resistance trained using previously assayed and publicly available *in vitro* mutant data. For each inhibitor, we present tenfold cross-validation performance measures of both classification as well as regression statistical learning algorithms. Two approaches are analyzed in each case, based on the use of either relative frequencies or counts of n-grams to represent mutant protein sequences as feature vectors. To the best of our knowledge, this is the first reported study on predictive models of HIV-1 PR and RT drug resistance developed by implementing n-grams to generate sequence attributes. Our technique is complementary to other sequence-based approaches and is competitive in performance. In a novel application, we classify every pair of RT inhibitors as either potentially effective as part of a larger drug cocktail or a combination that should not be concomitantly administered, with results that closely mirror available clinical and experimental data.**

Keywords-HIV; antiviral therapy; genotype-phenotype correlation; mutagenesis; machine learning

I. INTRODUCTION

Since the introduction of highly active antiretroviral therapy (HAART) in the mid-1990's for treating individuals infected with human immunodeficiency virus type 1 (HIV-1), developed countries around the world have observed substantial decreases in acquired immunodeficiency syndrome (AIDS) mortality rates, affording many the potential to live near-normal life spans with a chronic yet manageable condition [1]. The U.S. Food and Drug Administration (FDA) has approved a number of medications aimed at inhibiting essential viral proteins, with the HIV-1 protease (PR) and reverse transcriptase (RT) enzymes representing the two most common targets. Successful treatment for HIV-1 infection requires the co-administration of multiple drugs, and HAART refers to specific cocktails, many containing both PR and RT inhibitors, that are capable of effectively reducing viral replication to below detectable limits. Strict adherence to these treatment regimens is essential to preventing the development of PR or RT drug resistance mutations, whether due to polymorphisms attributable to the error-prone nature of DNA polymerization by RT or to nonpolymorphic treatment selected mutations.

While phenotype testing quantitatively measures the susceptibility of an HIV-1 PR or RT target to an inhibitor relative to that of a wild-type drug-sensitive control, significantly less expensive and faster genotype testing detects any target sequence mutations previously recognized to be associated with resistance to one or more drugs. Consequently, there have been growing demands for computational methods to rapidly and accurately predict phenotype from genotype. Several algorithmic methods relying solely on sequence-based features have been reported for predicting susceptibility to HIV-1 PR and RT inhibitors [2-4], including techniques that employ supervised classification and regression statistical machine learning tools [5-8], while most recent efforts in this arena are evaluating clinical data based on known patient outcomes as supplementary attributes [9, 10]. For one study in particular, Rhee *et al.* [11] systematically developed predictive sequence-based supervised classification and regression models of resistance with respect to each of 16 medications for treating HIV-1 infection: seven protease inhibitors (PIs), six nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs), and three nonnucleoside reverse transcriptase inhibitors (NNRTIs). Employing updated versions of the same sequence datasets, available online from the Stanford University HIV Drug Resistance Database [12], here we similarly develop predictive models of resistance to each of 19 drugs (eight PIs, eight NRTIs, and three NNRTIs) based on the use of n-grams to generate feature vector attributes for representing HIV-1 PR and RT sequences. The models yield performance measures that are comparable to those of Rhee *et al.*, and more generally, the methodology complements other sequence-based approaches. To the best of our knowledge, this is the first reported use of n-grams for developing predictive models of HIV-1 PR and RT drug resistance.

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

II. MATERIALS AND METHODS

A. Datasets

All HIV-1 PR and RT sequences used for this study, as well as their respective drug susceptibility values, are available from the Stanford University HIV Drug Resistance Database [12]. Sequences in the database correspond to both clinical and laboratory viral isolates, and we exclude those that display electrophoretic mixtures of two or more amino acids at any PR or RT sequence position as well as those that contain indels (i.e., insertions or deletions). The remaining sets of 548 PR and 331 RT sequences each consist of distinct mutational patterns defined exclusively by residue substitutions at one or more positions relative to the respective proteins based on the HIV-1 subtype B consensus wild-type sequence.

Names and abbreviations for the 19 drugs of this study, grouped by inhibitor class, are provided in Table I. Each PI is associated with its own subset of the 548 mutant HIV-1 PR sequences, corresponding to those for which susceptibility values with respect to the particular inhibitor are reported in the Stanford database; similarly, a distinct subset of the 331 mutant HIV-1 RT sequences is selected for each of the RT inhibitors (Table I, right column). Mutant drug resistance data are based on use of the PhenoSense assay (Monogram Biosciences, South San Francisco, CA), which reports susceptibility to an inhibitor as the ratio of 50% inhibitory concentration (IC_{50}) for the mutant relative to that for a drug-sensitive, wild-type control (i.e., fold change).

Following threshold fold change values and drug susceptibility classifications (sensitive/S, intermediate/I, or resistant/R) established by Rhee *et al.* [11], mutant HIV-1 PR or RT sequences comprising each inhibitor dataset are categorized as follows (Table I). For each of the PI datasets, mutant PR sequences with fold change values < 3 are sensitive, 3 – 20 are intermediate, and > 20 are resistant. For datasets corresponding to all three NNRTIs and the NRTIs AZT, 3TC, and FTC, mutant RT sequences with fold change values < 3 are sensitive, 3 – 25 are intermediate, and > 25 are resistant. For datasets corresponding to the NRTIs ddI, ddC, d4T, and TDF, mutant RT sequences with fold change values < 1.5 are sensitive, 1.5 – 3 are intermediate, and > 3 are resistant. For the dataset corresponding to the NRTI ABC, mutant RT sequences with fold change values < 2 are sensitive, 2 – 6 are intermediate, and > 6 are resistant. All fold changes are log-transformed and standardized prior to analysis.

B. Sequence Attributes Based on n-Grams

An n-gram (n-tuple) is a subsequence of n consecutive characters, and a sliding window of size n can be used to represent any sequence as an ordered set of n-grams. In addition to serving as a valuable tool for the statistical language-independent analysis of text [13], n-grams have been used to characterize protein sequences in various applications, including function prediction [14], secondary structure prediction [15], tertiary structure classification [16-18], and domain-domain interactions [19]. In this work we evaluate two methods for representing PR and RT protein sequences as attribute vectors, namely the relative frequency and the counts approaches, using n-grams of size $n = 2$ in both cases to directly compare methods. As will become apparent from the details provided below, feature vector dimensionality based on the counts approach becomes prohibitively large with any further increase to n-grams size.

TABLE I. DISTRIBUTION OF MUTANT HIV-1 SEQUENCE ISOLATES

Drug	Isolate Phenotypes (%) [a] S	I	R	Total
Protease Inhibitors				
Amprenavir (APV)	63	26	11	495
Atazanavir (ATV)	49	29	22	200
Indinavir (IDV)	53	26	21	502
Lopinavir (LPV)	46	22	32	320
Nelfinavir (NFV)	39	28	33	526
Ritonavir (RTV)	50	20	30	473
Saquinavir (SQV)	61	18	21	509
Tipranavir (TPV)	78	11	11	47
Nucleoside / Nucleotide RT Inhibitors				
Lamivudine (3TC)	29	18	53	244
Abacavir (ABC)	28	45	27	237
Zidovudine (AZT)	50	23	27	240
Stavudine (d4T)	53	36	11	242
Zalcitabine (ddC)	39	52	9	161
Didanosine (ddI)	51	43	6	243
Emtricitabine (FTC)	31	13	56	52
Tenofovir (TDF)	65	25	10	167
Nonnucleoside RT Inhibitors				
Delavirdine (DLV)	53	20	27	304
Efavirenz (EFV)	53	22	25	296
Nevirapine (NVP)	43	11	46	307

a. S, sensitive; I, intermediate; R, resistant; category thresholds provided in section II*A*.

Based on a 20-letter protein alphabet, there exist $20^n = 20^2 = 400$ distinct types of n-grams that are of size $n = 2$. For each inhibitor-specific dataset of protein sequences, the relative frequency approach is implemented as follows: 1) represent each sequence in the dataset as an ordered set of n-grams; 2) calculate the number of times that each of the 400 distinct types of n-grams appears collectively among all the sequences in the dataset (absolute frequencies); 3) divide the absolute frequencies by the total number of n-grams generated by all the sequences in the dataset (relative frequencies); 4) represent each sequence in the dataset as an ordered vector of the relative frequencies associated with its ordered set of n-grams. Since the HIV-1 PR and RT sequences are 99 and 543 residues in length, respectively, the relative frequency approach yields feature vectors for these sequences that are shorter by one in dimension.

With the counts approach on the other hand, each PR or RT sequence is represented by a 400-dimensional vector whose components each correspond to a particular type of n-gram, and whose value at each component is the absolute frequency of the respective n-gram based solely on number of occurrences in the given sequence. For $n = 3$ there are 8000 distinct types of n-grams, corresponding to a prohibitively large number of features relative to the number of sequences in each inhibitor dataset; hence, this study focuses only on n-grams of size $n = 2$.

C. Statistical Learning Algorithms and Performance

Each of the 19 inhibitor-specific sequence datasets is represented in four distinct ways, based on selections made for the type of input attributes used in feature vectors (relative frequency versus counts approaches) as well as measure of output drug susceptibilities (categorical versus numerical fold changes). Classification (regression) predictive models are trained with those datasets that contain categorical (numerical) outputs, respectively. Using the Weka software package [20], we implement the random forest (RF) classification [21] and the reduced-error pruned tree (REPTree) regression [20] statistical learning algorithms to derive models for classifying mutant sequences; in the latter case, classification is made based on where the predicted numerical fold change values fall relative to the designated category threshold values. We use software defaults for algorithm parameter values, with the following exceptions: RF is implemented using 100 trees (default is 10); and to reduce variance, a bootstrap aggregating (bagging) meta-learner [22] is applied with REPTree as the base learner.

We evaluate prediction performance of the algorithms on the datasets by using stratified tenfold cross-validation testing. For RF classification performance, we report the accuracy or success rate (i.e., proportion of correct predictions); the balanced error rate, given by

$$\mathrm{BER} = \frac{1}{3} \times \sum_{i=1}^{3} (1 - \mathrm{TPR}_i), \qquad (1)$$

where TPR_i are the category-specific true positive rates; the area under the receiver operating characteristic (ROC) curve, given by

$$\mathrm{AUC} = \sum_{i=1}^{3} \mathrm{AUC}_i \times \mathrm{p}_i, \qquad (2)$$

where the AUC_i are based on one-against-all classifications (i.e., category i sequences versus the remaining two categories combined), and the p_i are proportions of category i sequences in the dataset; the kappa statistic [23], a measure of performance that takes into account chance agreements between the actual and predicted sequence categories; and the out-of-bag (OOB) error rate. Decreasing values of BER and OOB toward 0, and increasing values of overall accuracy, AUC, and kappa toward 1, correspond to more accurate classifier predictions. For REPTree regression performance, we report the correlation coefficient (either r or r^2) between actual and predicted drug susceptibility values (i.e., the log-transformed and standardized fold changes) for the dataset sequences, the mean-squared error (mse), and the classification accuracy for the dataset sequences based on their predicted values. Finally, all statistical significance results make use of appropriate (paired, Student's, or Welch's) t-tests for the calculation of p-values.

III. Results and Discussion

A. Classification and Regression Summaries

Mean accuracy (2 types of n-grams datasets × 2 learning methods) is highest for the PIs (79.8%) as compared with the NRTIs (77.2%) and the NNRTIs (76.3%); however, of the three pairs of inhibitor classes, only the PIs and NNRTIs display a statistically significant mean accuracy difference ($p < 0.01$) (Table II). A comparison of our predictive accuracies in Table II with those reported by Rhee *et al.* (mean accuracy by class from their manuscript: PIs, 78.2%; NRTIs, 75.9%; and NNRTIs, 83.0%) [11] reveals statistically significant mean accuracy differences with respect to the PI ($p < 0.05$) and NNRTI ($p < 0.001$) inhibitor classes, which is reflective of our prediction advantage with PIs and that of Rhee *et al.* with NNRTIs. The inhibitors in Table II displaying the highest and lowest mean accuracy, respectively by class, are ritonavir (86.0%) and atazanavir (75.3%) for the PIs; emtricitabine (90.5%) and abacavir (67.3%) for the NRTIs; and nevirapine (80.5%) and delavirdine (73.3%) for the NNRTIs. Our results agree with those of Rhee *et al.* with respect to the PI and NNRTI classes; however, lamivudine achieved the highest mean accuracy among the NRTIs in their study, which did not include emtricitabine. With our approach, lamivudine achieved the second highest ranking in mean accuracy, surpassed only by emtricitabine among the NRTIs.

Averaging over all 19 inhibitors and both REPTree and RF learning methods, no statistically significant difference is observed between mean accuracies using the relative frequency (78.4%) and the counts (77.9%) approaches with n-grams for generating feature vectors to represent dataset sequences ($p = 0.17$); similarly, averaging over all 19 inhibitors as well as both the relative frequency and the counts approaches, no statistically significant difference is observed between mean accuracies using the REPTree (78.4%) and the RF (77.9%) learning methods ($p = 0.44$). By inhibitor class, REPTree regression (78.7%) displays higher mean accuracy than RF classification (74%, $p < 0.001$) for the NNRTIs, while no statistically significant differences in mean accuracies are evident for the PIs ($p = 0.08$) or the NRTIs ($p = 0.64$).

Additional evaluation metrics based on REPTree regression (r^2 and mse) and RF classification (OOB, BER, kappa, and AUC) are reported in an online appendix (http://proteins.gmu.edu/automute/HIV_Suppl.pdf), which consists of Supplementary Table S1 (relative frequency approach) and Supplementary Table S2 (counts approach). Averaging over all 19 inhibitors, the mean correlation coefficient associated with REPTree regression is the only performance measure for which there exists a statistically significant difference between the relative frequency (0.72) and the counts (0.70) approaches ($p < 0.02$). On the other hand, averaging over all 19 inhibitors as well as both the relative frequency and the counts approaches, we obtain the following results based on comparisons of mean values by inhibitor class: first, with respect to r^2 and

TABLE II. PREDICTIVE ACCURACY OF MODELS

Drug	Relative Frequency		Counts		Drug Mean
	REPTree	RF	REPTree	RF	
Protease Inhibitors					
APV	0.81	0.80	0.80	0.80	0.80
ATV	0.74	0.75	0.76	0.76	0.75
IDV	0.78	0.80	0.75	0.80	0.78
LPV	0.80	0.82	0.80	0.81	0.81
NFV	0.80	0.80	0.79	0.82	0.80
RTV	0.87	0.86	0.87	0.84	0.86
SQV	0.80	0.79	0.80	0.80	0.80
TPV	0.75	0.79	0.75	0.81	0.78
AVG	0.79	0.80	0.79	0.81	0.80
Nucleoside / Nucleotide RT Inhibitors					
3TC	0.89	0.87	0.87	0.90	0.88
ABC	0.68	0.68	0.66	0.67	0.67
AZT	0.75	0.75	0.73	0.70	0.73
d4T	0.74	0.79	0.76	0.78	0.77
ddC	0.80	0.75	0.80	0.76	0.78
ddI	0.69	0.73	0.69	0.71	0.71
FTC	0.96	0.83	0.94	0.89	0.91
TDF	0.75	0.75	0.68	0.74	0.73
AVG	0.78	0.77	0.77	0.77	0.77
Nonnucleoside RT Inhibitors					
DLV	0.76	0.70	0.76	0.71	0.73
EFV	0.78	0.74	0.76	0.73	0.75
NVP	0.84	0.79	0.82	0.77	0.81
AVG	0.79	0.74	0.78	0.74	0.76

kappa, no statistically significant mean differences are observed between any pair of inhibitor classes; next, with respect to mse, statistically significant mean differences are observed with every inhibitor class pair ($p < 0.001$ in all three cases); third, with respect to OOB and AUC, statistically significant mean differences are observed only between the PIs and each of the other two classes of HIV-1 RT inhibitors ($p < 0.02$ in both cases); and lastly, with respect to BER, a statistically significant mean difference is observed only between PIs and NNRTIs ($p < 0.02$).

B. Sequence Contribution to n-Grams Prediction

In this section we focus exclusively on the relative frequency approach for generating sequence feature vectors. Recall that the vector components (i.e., the input attributes) correspond to ordered consecutive pairs of amino acid sequence positions obtained using a sliding window of size $n = 2$. For example, since HIV-1 PR sequences are 99 residues long, every PI dataset consists of PR sequences represented as 98-dimensional relative frequency vectors, where the value at vector attribute i represents the pair of positions $(i, i + 1)$ in the sequence.

For each inhibitor dataset, a relatively small number of the vector input attributes are used to represent the tree nodes of any REPTree regression model, where the most information-rich attribute is selected as a root node. Briefly, for each attribute an optimal threshold value is computed to sort the sequences, one that results in the most divergent drug susceptibility (i.e., fold change) values between both sequence subsets; the attribute that displays maximal degree of subset divergence is the most information-rich and is selected as the root node. Starting from the root, a tree consists of multiple levels of nodes connected by branches, where two branches (representing either side of the threshold value for the attribute selected at the node) extend from each node to the next tree level. The procedure described for the root node (level 0) is repeated to select input attributes for nodes at subsequent levels of the tree, with more informative attributes being associated with nodes at increasingly higher tree levels approaching the root.

TABLE III. ATTRIBUTES SELECTED VIA REPTREE REGRESSION

Drugs	Root Node [a]	Level 1 Nodes [a]	Level 2 Nodes [a]
PIs			
APV	10	84, **87**	32, **34**, 53
ATV	54	73	32, 50
IDV	54	45, 53	72, 83, 90
LPV	54	45	77, 84
NFV	10	**54**, 87	29, **75**, 83, 90
RTV	54	9, 84	19, 82, 84
SQV	70	10, 83	47, 54, 90
TPV	90	**52**, 56	40, 73
NRTIs			
3TC	183	64	66
ABC	183	115, 214	64, 101, 114, 118
AZT	67	166, 210	76, 214
d4T	209	76, 177	66, 67
ddC	115	134, 183	65, 117
ddI	150	**43**, 61	39, **183**
FTC	183	123, 214	40
TDF	214	34, 65	**68**, **227**, 285
NNRTIs			
DLV	102	165, 180	69, 100, 190, 209
EFV	102	189	99, 188
NVP	189	103, 172	173, 180

a. Regular font, both IAS and TSM sets of positions; bold, TSM only; underlined, neither.

The REPTree regression model performance results of the previous section involved bagging multiple trees (10 iterates by default) for each inhibitor dataset. With a focus on the tree associated with the first iterate in each case, Table III summarizes the input attributes (i.e., feature vector component numbers) selected for the root node as well as for nodes at the next two levels. We compare the HIV-1 PR or RT sequence positions corresponding to these input attributes with the following two groups of inhibitor-specific subsets of sequence positions: those at which residue substitutions occur that are associated with drug resistance, according to an expert panel (International Antiviral Society – USA, IAS) [24, 25]; and those at which residue substitutions occur that are significantly more common in treated versus untreated individuals (nonpolymorphic treatment-selected mutations, TSM) [26]. As indicated in Table III, each of the 19 root node attributes (i) corresponds to a sequence position (i or $i + 1$) that appears in the respective inhibitor-specific subsets of both the IAS and TSM groups, while the majority (> 75%) of attributes at the next two levels of nodes correspond to

TABLE IV. REPTREE REGRESSION CORRELATION COEFFICIENTS USING THE RELATIVE FREQUENCY METHOD

	NRTIs								NNRTIs		
Train / Test	*3TC*	*ABC*	*AZT*	*d4T*	*ddC*	*ddI*	*FTC*	*TDF*	*DLV*	*EFV*	*NVP*
NRTIs											
3TC	0.98	0.69	-0.08	0.01	0.45	0.38	0.99	-0.31	-0.13	-0.17	-0.25
ABC	0.85	0.91	0.29	0.42	0.62	0.63	0.93	0.05	-0.10	-0.05	-0.14
AZT	0.11	0.44	0.91	0.78	0.27	0.35	0.32	0.60	-0.07	-0.01	-0.05
d4T	0.18	0.51	0.79	0.91	0.57	0.58	0.29	0.53	-0.07	-0.02	-0.06
ddC	0.57	0.63	0.16	0.47	0.90	0.79	0.08	-0.07	-0.13	-0.17	-0.22
ddI	0.45	0.68	0.21	0.56	0.86	0.91	0.84	0.03	-0.10	-0.07	-0.13
FTC	0.94	0.69	0.03	0.05	0.41	0.36	1.00	-0.27	-0.13	-0.17	-0.24
TDF	-0.42	-0.06	0.68	0.48	-0.19	-0.05	-0.34	0.82	0.04	0.11	0.10
NNRTIs											
DLV	-0.14	-0.15	0.02	-0.03	-0.10	-0.10	-0.20	-0.07	0.87	0.51	0.60
EFV	-0.13	-0.02	0.14	0.05	-0.10	-0.06	-0.13	0.06	0.55	0.91	0.72
NVP	-0.10	-0.01	0.15	0.09	-0.13	-0.06	-0.11	0.02	0.60	0.73	0.92

positions that either overlap both groups or appear exclusively in the TSM subsets.

C. Predicting Outcomes for Co-Administered Pairs of HIV-1 RT Drugs

The effectiveness of recommended HAART drug cocktails (Antiretroviral Guidelines for Adults and Adolescents, U.S. Department of Health and Human Services [27]) in controlling HIV-1 replication stems from the fact that they either simultaneously inhibit more than one target (e.g., one PI and two NRTIs), or as with certain combinations of RT inhibitors, they target the same protein with drugs that possess non-overlapping mutational patterns of cross-resistance (e.g., one NNRTI and two NRTIs). In addition to the potential for cross-resistance, serious adverse events such as antagonism, toxicity, and reduced efficacy exclude the concomitant use of certain drug pairs. Here we implement a method to predict outcomes associated with co-administering every pair of RT inhibitors, results that compare favorably with available clinical and experimental data.

Based on the relative frequency approach for representing RT sequences as feature vectors, we select one of the previously trained, inhibitor-specific, bagged REPTree regression models and use it to predict the fold changes for another inhibitor-specific dataset of sequences as a test set. The correlation coefficients ($-1 \leq r \leq 1$) between the actual and predicted fold change values associated with all possible training/testing pairs of inhibitor datasets are presented in Table IV, where values down the main diagonal reflect the resubstitution errors corresponding to using the same dataset for both training and testing. High positive correlations are expected between pairs of inhibitors that display similar susceptibility profiles with respect to RT sequence mutational patterns (i.e., cross-resistance), as well as those inhibitors that, when co-administered, are more likely to lead to adverse events; conversely, drugs displaying incongruent susceptibility profiles are often effective when prescribed together and are expected to generate near zero or negative correlations.

Shaded regions in Table IV, where the vast majority of correlation coefficients are slightly negative (remainder satisfy $r < 0.15$), reflect the effectiveness of HAART regimens exclusively targeting RT that incorporate drugs from both NNRTI and NRTI classes. Current guidelines for selection of two NRTIs to include in these cocktails suggest (FTC or 3TC) / TDF (preferred), (FTC or 3TC) / AZT (acceptable), and (FTC or 3TC) / d4T (alternative to AZT-based) [27], with all corresponding training/testing pairs in Table IV displaying relatively low or negative correlations in line with expectations. Though the combinations (FTC or 3TC) / ABC (alternative to TDF-based) are also occasionally prescribed, their continued administration is often precluded by adverse conditions that include hypersensitivity reactions in the presence of the HLA-B*5701 allele [28, 29], cardiovascular events [30, 31], and inferior virologic response in patients with high viral loads [32]; hence, consistent with our expectations, the corresponding training/testing pairs all have relatively high correlations. Similarly high correlations are reported in Table IV for NRTI pairs that current guidelines explicitly recommend avoiding: AZT / d4T (antagonism), FTC / 3TC (similar resistance profiles, non-additive antiviral activity), ddI / d4T (toxicity, inferior efficacy), and (FTC or 3TC) / ddI (inferior efficacy) [27]. Lastly, due to inferior efficacy in a large-scale clinical trial [33], current guidelines recommend drug cocktails contain at most one NNRTI [27]. Consistent with these results, we report high correlations in Table IV for all NNRTI pairs.

IV. CONCLUSIONS

In this study, n-grams are used for the first time to represent HIV-1 PR and RT genotypes as feature vectors in order to develop predictive models of drug resistance. The approach complements other sequence-based methods

and performs equally well. An investigation into feature selections made by tree-based models reveals how predictive accuracy is overwhelmingly based on inhibitor-specific subsets of the PR and RT sequence positions at which residue substitutions occur that are known to be either associated with drug resistance or selected for by the treatment. In an innovative application of the RT inhibitor regression models, we describe a method to accurately predict every pair of RT drugs as either effective together as part of a multi-drug cocktail, or a combination that should not be co-administered.

ACKNOWLEDGMENT

We are grateful to researchers affiliated with the Stanford University HIV Drug Resistance Database for providing online access to genotype-phenotype correlation data characterizing HIV-1 PR and RT sequences.

REFERENCES

[1] S. Broder, "The development of antiretroviral therapy and its impact on the HIV-1/AIDS pandemic," Antiviral Res, vol. 85, 2010, pp. 1-18.

[2] A. G. DiRienzo, V. DeGruttola, B. Larder, and K. Hertogs, "Non-parametric methods to predict HIV drug susceptibility phenotype from genotype," Stat Med, vol. 22, 2003, pp. 2785-2798.

[3] E. Puchhammer-Stockl, C. Steininger, E. Geringer, and F. X. Heinz, "Comparison of virtual phenotype and HIV-SEQ program (Stanford) interpretation for predicting drug resistance of HIV strains," HIV Med, vol. 3, 2002, pp. 200-206.

[4] M. Zazzi, L. Romano, G. Venturi, R. W. Shafer, C. Reid, et al., "Comparative evaluation of three computerized algorithms for prediction of antiretroviral susceptibility from HIV type 1 genotype," J Antimicrob Chemother, vol. 53, 2004, pp. 356-360.

[5] N. Beerenwinkel, M. Daumer, M. Oette, K. Korn, D. Hoffmann, et al., "Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes," Nucleic Acids Res, vol. 31, 2003, pp. 3850-3855.

[6] S. Draghici and R. B. Potter, "Predicting HIV drug resistance with neural networks," Bioinformatics, vol. 19, 2003, pp. 98-107.

[7] D. Wang and B. Larder, "Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks," J Infect Dis, vol. 188, 2003, pp. 653-660.

[8] K. Wang, E. Jenwitheesuk, R. Samudrala, and J. E. Mittler, "Simple linear model provides highly accurate genotypic predictions of HIV-1 drug resistance," Antivir Ther, vol. 9, 2004, pp. 343-352.

[9] M. C. Prosperi, M. Rosen-Zvi, A. Altmann, M. Zazzi, S. Di Giambenedetto, et al., "Antiretroviral therapy optimisation without genotype resistance testing: a perspective on treatment history based models," PLoS One, vol. 5, 2010, pp. e13753.

[10] M. Zazzi, F. Incardona, M. Rosen-Zvi, M. Prosperi, T. Lengauer, et al., "Predicting response to antiretroviral treatment by machine learning: the EuResist project," Intervirology, vol. 55, 2012, pp. 123-127.

[11] S. Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, et al., "Genotypic predictors of human immunodeficiency virus type 1 drug resistance," Proc Natl Acad Sci U S A, vol. 103, 2006, pp. 17355-17360.

[12] http://hivdb.stanford.edu/.

[13] M. Damashek, "Gauging Similarity with n-Grams: Language-Independent Categorization of Text," Science, vol. 267, 1995, pp. 843-848.

[14] Q. Dong, S. Zhou, L. Deng, and J. Guan, "Gene ontology-based protein function prediction by using sequence composition information," Protein Pept Lett, vol. 17, 2010, pp. 789-795.

[15] J. K. Vries, X. Liu, and I. Bahar, "The relationship between n-gram patterns and protein secondary structure," Proteins, vol. 68, 2007, pp. 830-838.

[16] B. Y. Cheng, J. G. Carbonell, and J. Klein-Seetharaman, "Protein classification based on text document classification techniques," Proteins, vol. 58, 2005, pp. 955-970.

[17] E. G. Mansoori, M. J. Zolghadri, and S. D. Katebi, "Protein superfamily classification using fuzzy rule-based classifier," IEEE Trans Nanobioscience, vol. 8, 2009, pp. 92-99.

[18] C. H. Wu, S. Zhao, H. L. Chen, C. J. Lo, and J. McLarty, "Motif identification neural design for rapid and sensitive protein family search," Comput Appl Biosci, vol. 12, 1996, pp. 109-118.

[19] K. X. Zhang and B. F. Ouellette, "GAIA: a gram-based interaction analysis tool--an approach for identifying interacting domains in yeast," BMC Bioinformatics, vol. 10 Suppl 1, 2009, pp. S60.

[20] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using Weka," Bioinformatics, vol. 20, 2004, pp. 2479-2481.

[21] L. Breiman, "Random forests," Machine Learning, vol. 45, 2001, pp. 5-32.

[22] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, 1996, pp. 123-140.

[23] Y. Bishop, S. Fienberg, and P. Holland, Discrete Multivariate Analysis - Theory and Practice. Cambridge, MA: MIT Press, 1975.

[24] V. A. Johnson, F. Brun-Vezinet, B. Clotet, B. Conway, D. R. Kuritzkes, et al., "Update of the drug resistance mutations in HIV-1: 2005," Top HIV Med, vol. 13, 2005, pp. 51-57.

[25] V. A. Johnson, F. Brun-Vezinet, B. Clotet, H. F. Gunthard, D. R. Kuritzkes, et al., "Update of the drug resistance mutations in HIV-1: December 2010," Top HIV Med, vol. 18, 2010, pp. 156-163.

[26] S. Y. Rhee, W. J. Fessel, A. R. Zolopa, L. Hurley, T. Liu, et al., "HIV-1 Protease and reverse-transcriptase mutations: correlations with antiretroviral therapy in subtype B isolates and implications for drug-resistance surveillance," J Infect Dis, vol. 192, 2005, pp. 456-465.

[27] http://www.aidsinfo.nih.gov/ContentFiles/AdultandAdolescentGL.pdf.

[28] S. Mallal, E. Phillips, G. Carosi, J. M. Molina, C. Workman, et al., "HLA-B*5701 screening for hypersensitivity to abacavir," N Engl J Med, vol. 358, 2008, pp. 568-579.

[29] M. Saag, R. Balu, E. Phillips, P. Brachman, C. Martorell, et al., "High sensitivity of human leukocyte antigen-b*5701 as a marker for immunologically confirmed abacavir hypersensitivity in white and black patients," Clin Infect Dis, vol. 46, 2008, pp. 1111-1118.

[30] C. A. Sabin, S. W. Worm, R. Weber, P. Reiss, W. El-Sadr, et al., "Use of nucleoside reverse transcriptase inhibitors and risk of myocardial infarction in HIV-infected patients enrolled in the D:A:D study: a multi-cohort collaboration," Lancet, vol. 371, 2008, pp. 1417-1426.

[31] S. W. Worm, C. Sabin, R. Weber, P. Reiss, W. El-Sadr, et al., "Risk of myocardial infarction in patients with HIV infection exposed to specific individual antiretroviral drugs from the 3 major drug classes: the data collection on adverse events of anti-HIV drugs (D:A:D) study," J Infect Dis, vol. 201, 2010, pp. 318-330.

[32] P. E. Sax, C. Tierney, A. C. Collier, M. A. Fischl, K. Mollan, et al., "Abacavir-lamivudine versus tenofovir-emtricitabine for initial HIV-1 therapy," N Engl J Med, vol. 361, 2009, pp. 2230-2240.

[33] F. van Leth, P. Phanuphak, K. Ruxrungtham, E. Baraldi, S. Miller, et al., "Comparison of first-line antiretroviral therapy with regimens including nevirapine, efavirenz, or both drugs, plus stavudine and lamivudine: a randomised open-label trial, the 2NN Study," Lancet, vol. 363, 2004, pp. 1253-1263.