JEAP

ISSN: 2525-815X

Journal of Environmental Analysis and Progress

Journal homepage: www.jeap.ufrpe.br/

10.24221/jeap.3.3.2018.2034.319-329

Features in HIV genotypes associated with failure in the computational prediction of patients' response to antiretroviral treatment

Rogério S. Rosa[a], Ádamo Y. B. da Silva[b], Viviane M. S. de Morais[c], Rafael H. S. Santos[d], Katia S. Guimarães[e]

[a] Centro de Tecnologias Estratégicas do Nordeste-CETENE. Av. Prof. Luís Freire, n. 01, Recife, Pernambuco, Brasil, CEP: 50740-540. E-mail: rogerio.bioinfo@gmail.com.
[b] Universidade Federal de Pernambuco-UFPE, Centro de Biociências, Recife, Brasil. CEP: 50730-120.
[c] Universidade Federal de Pernambuco-UFPE, LIKA, Recife, Brasil. CEP: 50670-901.
[d] Technical University of Madrid, School of Computer Engineering, Madrid, Spain.
[e], UFPE, Centro de Informática. Av. Jornalista Aníbal Fernandes, s/n, Recife, Brasil. CEP: 50740-560.

ARTICLE INFO

Received 03 Jul 2018
Accepted 18 Jul 2018
Published 31 Jul 2018

ABSTRACT

HIV acts by attacking the immune system and gradually destroying the TCD4+ defense cells. Without adequate treatment, the carriers develop the most severe form of the infection, AIDS, when the patient can be afflicted by opportunistic diseases that inevitably lead to death. Fortunately, with the advent of the highly active antiretroviral therapy (HAART), the mortality of people with HIV is decreasing. However, mutations can occur in the genotype of the virus, generating drug-resistant phenotypes. Computational methods have been used to predict whether a given strain is drug-resistant, and to which drugs this resistance occurs, thereby increasing the chances of success of the prescribed treatment regimen. However, these methods are not always accurate in their task. In this context, by applying Feature Selection methods and estimating Decision Tree models, we investigated patterns in Protease and Reverse Transcriptase enzyme sequences, as well as in patients' clinical data, which can lead to correct or incorrect computational prediction. As a result, we identified 21 features that are highly informative, 11 which tend to lead the methods to error, and eight that present both behaviors simultaneously, being able to predict the patient's response to therapy and at the same time may lead the predictor's methods to failure.

Keywords: HIV, antiretroviral therapy, drug-resistance, phenotype, computational methods.

Introduction

The Human Immunodeficiency Virus (HIV) infection is one of the largest pandemics ever recorded in the history of health sciences (Defo, Kouotou & Richie, 2017; Egbe et al., 2017). An estimation performed in 2015 by the Joint United Nations Program on HIV/AIDS (UNAIDS) revealed that approximately 33.7 million people are living with the virus, and that, only in 2015, 2.1 million new infections were recorded. Since its discovery, the HIV infection has killed more than 36 million people. Despite these numbers, mortality from infection has declined considerably due to the new therapies employed to control the spread and progression of the infection to its most severe form. For example, in 2005 2 million AIDS-related deaths were recorded, while in 2015 this number dropped to approximately half (1.1 million people), revealing the importance of the efforts of health teams and researchers in the fight against the HIV (Avert Fundation, 2017).

The use of antiretrovirals (ARVs) has been of high relevance for the decay in the mortality rate of people living with HIV/AIDS (PLWHA). However, the use of these drugs showed another aspect of the infection: the emergence of mutations in the viral genome, which confers HIV resistance and consequently lead to treatment failure. Several strategies have been proposed to optimize PLWHA treatment to avoid therapeutic failures. Among these strategies, computational models have played a crucial role in establishing the link between the existence of HIV resistance mutations and the occurrence of therapy failure (Larder et al., 2007).

These mutations were evidenced using the first ARVs in monotherapy, which showed a natural

selection of resistant viral strains within a few months of treatment. This problem has been solved mainly with the use of highly active antiretroviral therapy (HAART), which attacks the virus with a combination of several types of antiretrovirals, acting at different sites of different viral replication proteins. Thus, for a viral strain to replicate, it must be resistant to all drugs used by the patient. Allied to HAART, several forms of prevention, diagnosis, treatment, and control have been studied to increase the chances of success for the newly diagnosed with the infection (May et al., 2014; Wang et al., 2009).

HIV genotyping has been used as the main source of information on resistance mutations. It consists in finding alterations on viral RNA that result in resistant phenotypes. It is the most common method of assessing HAART failure. This interpretation influences the choice of the treatment scheme that will be prescribed for the patient. Also, the use of the genetic sequence of the virus has been gaining prominence by improving the performance of computational models of prediction of response to therapy, becoming a parameter of extreme importance to guide the choice of new treatment schemes (Ozahata et al., 2015).

Computational methods play a key role in this process since they can quickly recognize patterns of viral genome mutations that are more frequent in resistant phenotypes already known and can also estimate the probability of each ARV scheme being used. Online platforms provide systems that, using computational prediction methods applied to databases updated continuously by research centers and partners, evaluate both HIV genetic data and historical patient data, such as previously used ARVs, HAART in use when therapy failed, history of clinical manifestations, TCD4+ cells (cells/dL) count and Viral Load (VL) (copies.μL^{-1}) (Revell et al., 2013).

The platforms use Machine Learning Methods (MLM), which work based on the concept of classification: the separation of entries based on a pattern present in the database. These computational methods have gained space in diagnostic and treatment centers, helping physicians in the choice of the best therapeutic scheme, especially in cases of high risk of therapeutic failure and in severe cases, where the choice of the appropriate antiretroviral may be the difference between life and death (Beerenwinkel et al., 2005).

Although computational prediction methods are efficient, they are not always accurate in determining the PLWHA response to the treatment. For this reason, an investigation of the causes that can lead predictions to fail is essential to increase the chances of defining more effective therapeutic strategies.

In a previous study (Rosa et al., 2014), we estimated and evaluated the performance of three different MLMs for this specific prediction task; they were: Multi-Layer Perceptron (MLP), Radial Base Functions (RBF), and Vector Support Machines (SVM). We showed statistical evidence that SVM is the most accurate method for the task, and, through a non-exhaustive data-mining procedure, we found codon rt184 to be a strong error indicator.

In this study, we apply exhaustively feature selection and another MLM called Decision Trees to identify, in the results of our previous work, patterns that lead the computational approaches to fail. We analyze the contribution of Protease (PR), Reverse Transcriptase (RT), TCD4+ count, and VL in the prediction process of the patient's response to antiretroviral therapy in each approach. As a result, we have identified 21 features that are directly linked to correct predictions, 11 that are linked to incorrect predictions, and eight features that are ambiguous (although they are reliable predictors, they are also strong candidates to induce MLM to failure, because they are highly variant codons). Decision trees show us that informative features are not necessarily correlated among them, and they can interact with poor predictors leading to better results.

Material and Methods

The pipeline developed for the analysis consists of three stages: Data Preprocessing, Feature Selection, and Decision Trees fitting. For those tasks, we used Weka and R Statistics (Hall et al., 2009; R Development Core Team, 2008).

Data preprocessing

The database used in our experiments was initially collected from the HIV Resistance Drug Database (Rhee et al., 2003). It contains data from 1,000 patients consisting of: (1) An identification number (ID), (2) VL values in $\log_{10}$, (3) TCD4+ cells count at the beginning of the treatment, (4) The RT and PR RNA sequences of the virus (99 codons of PR and 300 of RT), and (5) Patient's characteristic (class), representing the success of the antiretroviral scheme, measured after 16 weeks of treatment, as *responder* (if there was a decrease in VL value of at least 1 $\log_{10}$) or *non-responder* (otherwise).

In our analyses, we consider the raw data generated from our previous work (Rosa et al., 2014), in which we applied the classification methods MLP, RBF, and SVM. We considered eight datasets, as described in Table 1. Each dataset

contains all the features (VL, TCD4+ cells count, 99 PR codons and 300 RT codons) and one dichotomous variable (class). The classes differ in each data set according to the purpose of the analysis that will be employed in each of them. Original_Set contains the original data classified in responder or non-responder. EasyHard_Set patients are classified as *easy* (when all three classification methods correctly predict the patient's response to treatment) and *hard* (when all three classification methods fail the prediction). MLP_OutPut, RBF_OutPut, and SVM_OutPut contain the instances classified in *responder* or *non-responder* according to the output of each method. MLP_Map, RBF_Map, and SVM_Map contain the error/hit map for each approach, and each patient is classified as *correct* or *incorrect* (if the patient's response to treatment was predicted correctly or incorrectly). Except for EasyHard_Set, all datasets originally contained 1,000 patients.

Table 1. Number of samples by class in each dataset.

Dataset	Class	Original	SMOTE
Original_Set	non-responders	794	794
	responders	206	824
	Total	1,000	1,618
EasyHard_Set	easy	646	646
	hard	69	621
	Total	715	1,267
MLP_OutPut	non-responders	661	661
	responders	339	678
	Total	1,000	1,339
RBF_OutPut	non-responders	699	699
	Responders	301	752
	Total	1,000	1,451
SVM_OutPut	non-responders	720	720
	responders	280	700
	Total	1,000	1,420
MLP_Map	correct	755	755
	incorrect	245	735
	Total	1,000	1,490
RBF_Map	correct	807	807
	incorrect	193	772
	Total	1,000	1,579
SVM_Map	correct	844	844
	incorrect	156	780
	Total	1,000	1,624

The unbalance between classes of patients is a problem for the construction of decision trees. For example, in Original_Set there are 206 responders and 794 non-responders. This disequilibrium between classes may lead the computational approach to tend to the class containing the biggest number of samples. Hence it is necessary to resort to computational strategies to balance the number of instances between classes. We applied Synthetic Oversampling Minority Technique (SMOTE) (CHAWLA et al., 2002) to create "synthetic patients," and thus balance the classes in each dataset.

SMOTE was chosen because it is an algorithm that keeps the minority class homogeneous regarding attributes, as it traverses the data set and constructs the new samples comparing attribute by attribute among the closest k-neighbors to the right and left, randomly selecting a value for the attribute of the new feature. SMOTE is advantageous because it maintains a certain homogeneity between the original features and the synthetic features, given that it uses the original attributes to construct the new ones.

Feature selection

Considering all the attributes in the dataset, only a subset presents direct correlation with the classes. Determining this subset is a problem known as Feature Selection. Selecting the most suitable subset helps to avoid overfitting and can also improve the model performance, besides decreasing data processing time (Kumar & Minz, 2014).

We apply Weka's Method *InfoGainAttributeEval* to measure the information gain of each attribute about the class. As a result, we obtained an ordered list of features, from the highest information gain to the lowest. All features with information gain value greater than a given threshold (0.02) were selected. This process was applied to each one of the eight datasets.

Decision trees fitting

Decision Trees are constructed by choosing the main attribute, which labels the root node, and then placing elements under one of the two branching subtrees, according to a classification rule, and repeating that process until all the entries are separated. The attributes that are at the base of the tree are called leaves, and those that are between the leaves and the root node are the intermediate nodes.

This technique uses the simplest types of classification algorithms in Machine Learning. It is because interactions are represented in the classification model and that the observer can quickly evaluate and understand all stages of processing (which means that it does not suffer from the "black box" effect of Neural Networks, for example).

For the identification of the attributes that were more relevant for the task of prediction, features simultaneously selected by prediction of treatment response (Original_Set, MLP_Output, RBF_Output, and SVM_Output) were analyzed. Features related to the prediction of MLM failure in different datasets (MLP_Map, RBF_Map, and SVM_Map) were also considered simultaneously.

We fitted decision trees for each of the datasets (considering actual and synthetic patients). For this, we considered the subset of attributes resulting from the Feature Selection. We applied the Weka's implementation of Algorithm J48 for tree construction. We used 10-fold cross-validation and we used the following parameters to achieve better accuracy: binarySplits=True, confidenceFactor=0.1, minNumObj=20, useLaplace=True.

Results

Selected features

The features in each dataset were ranked according to their information gain (the full rank is not shown due to space limitation). TCD4 + count presented lower information gain in the eight datasets, which reveals that the counting of this type of cell at the beginning of the treatment is not informative for the prediction of the patient's response to antiretroviral treatment.

Table 2. Number of selected features by the dataset.

Dataset	Features
Original_Set	117
EasyHard_Set	128
MLP_OutPut	59
RBF_OutPut	89
SVM_OutPut	70
MLP_Map	83
RBF_Map	91
SVM_Map	77
Average	**86**
Standard Deviation	**17.06**

The number of features selected in each dataset is shown in Table 2. The largest number of features selected occurred in the Original_Set (117) dataset, while the smallest amount was in MLP_OutPut (59). On average, 86 features presented information gain about their respective classes. An interesting observation is that more variables were selected for the error map of each MLM than for generating its original output: MLP (output 59 / error map 83), RBF (output 89 / error map 91) and SVM (output 70 / error map 77). This finding suggests that identifying the patterns that lead MLM to fail in predicting patient response to treatment is a more complex task than classifying them into responder or non-responder.

The features simultaneously selected for Original_Set, MLP_OutPut, RBF_OutPut and SVM_OutPut were defined as *powerful predictors* because if they are relevant for the prediction in all MLMs, they are highly informative attributes. Formally, the feature set selected for a given dataset is denoted by *fs(dataset)*, then: (fs(Original_Set) ∩ fs(MLP_OutPut) ∩ fs(RBF_OutPut) ∩ fs(SVM_OutPut) = powerful_predictors.

The features simultaneously selected for error maps are those that present evidence of being strong candidates for inducing MLM to fail and were defined as *error predictors*, thus: (fs(MLP_Map) ∩ fs(RBF_Map) ∩ fs(SVM_Map) = error_predictors.

The intersection between powerful_predictors and error_predictors results in *ambiguous predictors*, which are those features with high predictive power, but which in certain scenarios may lead MLM to fail. Formally: (powerful_predictors ∩ error_predictors) = ambiguous_predictors.

Figure 1 shows a Venn Diagram for the features powerful_predictors and error_predictors. The top set contains the powerful_predictors, the bottom set contains the error_predictors, and the intersection of both contains the so-called ambiguous_predictors.

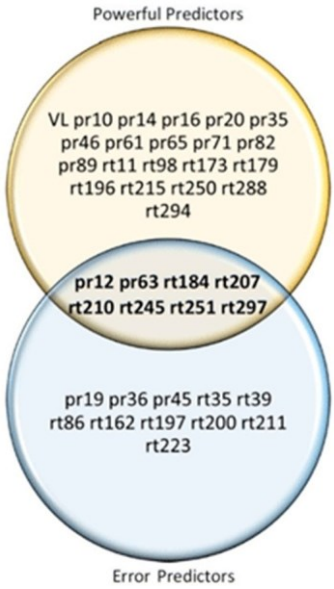

Figure 1. Venn diagram of powerful predictors and error predictors.

According to the results shown in Figure 1, considering the intersection zone, two codons of Protease (pr12 and pr63) and six codons of Reverse Transcriptase (rt184, rt207, rt210, rt245, rt251, and rt297) were identified as *ambiguous predictors*

(ambiguous_predictor). Such codons are highly informative but also induce MLM to error. We found an intersection between fs(EasyHard_Set) and ambiguous_predictors, and we observed that every feature in ambiguous_predictors also occurs in fs(EasyHard_Set), in other words, ambiguous_predictor ⊂ fs(EasyHard_Set).

Considering only the variables above the intersection zone, it can be identified 21 features: VL measured at beginning of treatment, codons of Protease enzyme (pr10, pr14, pr16, pr20, pr35, pr46, pr61, pr65, pr71, pr82 and pr89) and codons of the Reverse Transcriptase (rt11, rt98, rt173, rt179, rt196, rt215, rt288, rt250, and rt294). These were the most powerful features to predict patient response to treatment, and none of them induced MLM to fail. Thus, we named this set *only powerful predictors*: only_powerful_predictors = (powerful_predictors = ambiguous_preditors).

Analyzing the part below the intersection zone, 11 features were identified: three Protease codons (pr19, pr36 and pr45) and eight Reverse Transcriptase codons (rt35, rt39, rt86, rt162, rt197, rt200, rt211 and rt223), which induced MLM to fail and were not informative for predicting patient response to treatment. Therefore, we have the set *only error predictors*: (error_predictors − ambiguous_predictors) = only_error_predictors.

Decision trees

Accuracy of the models

Table 3 shows quality assessment indicators for each one of the eight decision tree models fitted. The results show that the models were able to identify the patterns in each of the datasets in a way that there was no absolute deviation tendency for any given class in any of the models (the accuracy was similar for both classes of each dataset). This result gives greater certainty about the performance of the models in predicting each class correctly.

The model with the highest accuracy was the one fitted to the data set EasyHard_Set (90.5%), indicating that the Decision Tree Model fitted for the EasyHard_Set was able to precisely separate patients for which the three MLM methods would correctly or incorrectly predict the patients' responses to the treatment.

Table 3. Accuracy of decision trees. TPR = True Positive Rate, FPR = False Positive Rate, ROC = Area under ROC curve.

Dataset	Class	TPR	FPR	Accuracy	ROC
Original_Set	responder	.861	.166	.833	.885
	non-responder	.834	.139	.862	.885
	Average	**.847**	**.152**	**.848**	**.885**
EasyHard_Set	easy	.929	.122	.888	.918
	hard	.878	.071	.922	.918
	Average	**.903**	**.096**	**.905**	**.918**
MLP_Output	responder	.702	.298	.697	.739
	non-responder	.702	.298	.707	.739
	Average	**.702**	**.298**	**.702**	**.739**
RBF_Output	responder	.767	.286	.714	.786
	non-responder	.714	.233	.767	.786
	Average	**.739**	**.259**	**.740**	**.786**
SVM_Output	responder	.811	.359	.699	.769
	non-responder	.641	.189	.768	.769
	Average	**.727**	**.275**	**.733**	**.769**
MLP_Map	correct	.886	.309	.747	.802
	incorrect	.691	.114	.855	.802
	Average	**.790**	**.213**	**.800**	**.802**
RBF_Map	correct	.896	.244	.794	.835
	incorrect	.756	.104	.874	.835
	Average	**.828**	**.175**	**.833**	**.835**
SVM_Map	correct	.896	.187	.838	.863
	incorrect	.813	.104	.878	.863
	Average	**.856**	**.147**	**.857**	**.863**

The least accurate decision tree was the one estimated for the MLP_Output dataset (70.2%), suggesting that the lack in the precision of the MLP Method itself (as reported in ROSA et al., 2014), may affect the prediction of its outcome.

All methods of decision tree models presented a considerable global performance with

an area under the ROC curve above 0.7. The EasyHard_Set dataset obtained the highest value among all areas, reaching more than 91%, while the dataset MLP_OtuPut had the lowest value (0.739).

Structure of decision trees

A brief description of the fitted trees is shown in Table 4. Trees with an average of 30 nodes were obtained (standard deviation of 4.25). The largest one was obtained for the data set SVM_OutPut (39 nodes and 20 leaves), while the smallest was for the data set EasyHard_Set (21 nodes and 11 leaves).

The MLP_OutPut and RBF_OutPut trees obtained the same values for nodes and leaves (35 and 18 for the respective parameters), which also happened for MLP_Map and SVM_Map trees (29 nodes and 15 leaves for both datasets). The dataset RBF_Map generated a tree with 31 nodes and 16 leaves and, lastly, the tree of the Original_Set dataset presented 27 nodes and 14 leaves. Trees averaged 15 leaves (standard deviation of 2.12).

Table 4. Summary of decision trees fitted.

Dataset	nodes	leaves	root
Original_Set	27	14	VL
EasyHard_Set	21	11	rt245
MLP_OutPut	35	18	rt82
RBF_OutPut	35	18	pr54
SVM_OutPut	39	20	pr54
MLP_Map	29	15	rt98
RBF_Map	31	16	rt200
SVM_Map	29	15	rt245
Average	**30**	**15.5**	
Standard Deviation	**4.25**	**2.12**	

The root is the main node of the tree because it represents the feature considered as most informative by the model. The trees constructed for Original_Set, MLP_OutPut and MLP_Map presented as root the features VL, rt82, and rt98, respectively, which belong to the list of variables only_powerful_predictors.

The trees for EasyHard_Set and SVM_Map selected the ambiguous feature rt245 as root. For RBF_Map the root variable was rt200, which belongs to the list of features only_error_predictors. The fitted models for RBF_Output and SVM_Output built their respective trees from pr54, which did not appear in any of our previous attribute classifications.

The fitted decision trees are illustrated in Figure 2. Due to space limitation, we present the trees only up to the third level. The nodes are labeled with a feature, the edges with a decision rule, and the leaves with the class resulting from the decision rule.

Each leaf of the generated trees has 3 values: the value of the class and, in parentheses, the total number of instances classified according to that rule, followed by the number of samples classified incorrectly. For example, in the estimated tree for the Original_Set dataset, if VL ≤ 4.0 then the patient is classified as 0 (non-responder) and, of the 416 instances under this rule, 40 were classified incorrectly (patients responders classified as non-responders), so 376 cases were classified correctly (in this case the accuracy was 89.6\%). In case VL > 4.0, codon pr10 is consulted, if it contains an Isoleucine (I) the patient is classified as non-responder, but if pr10 contains another amino acid, then another rule is applied (data not shown).

Comparison of selected features and decision trees

As described in the previous section, features resulting from the feature selection process were grouped into five categories: only_powerful_predictors, only_error_predictors, ambiguous_predictors, powerful_predictors.

Then, decision trees were estimated considering all the variables selected by the feature selection method. The tree-building algorithm evaluates the information gain of each variable and correlates them, choosing the feature subset (among those already selected) that maximizes the information gain and that, applying pruning strategy, does not cause overfitting. After that, we compared the feature sets chosen by Algorithm J48 for the construction of each tree to the feature sets which were classified in five categories in the first step.

Figure 3 presents six Venn diagrams. Each of them is composed of three feature sets: powerful_predictors, error_predictors, and the features existing in each decision tree model (denoted as *DT(dataset)*).

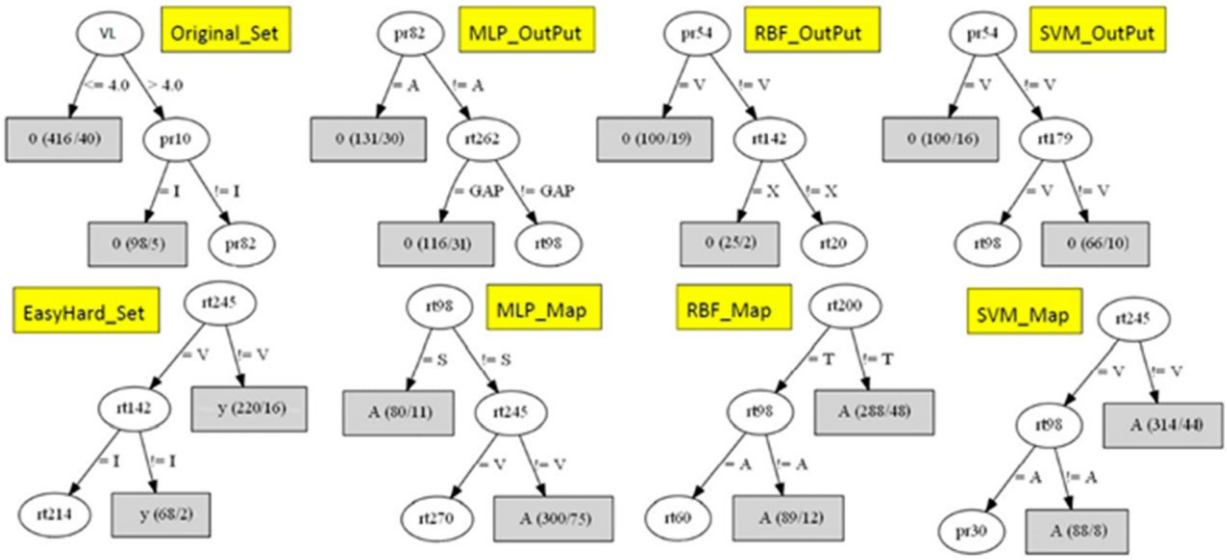

Figure 2. Partial illustration of the fitted decision trees. On leaf nodes: "0" denotes non-responders, "1" responders, "A" correct, "E" incorrect, "y" easy and "h" hard.

According to the results, enzyme codons rt98 and rt179 were the only only_powerful_predictors codons appearing in the six trees: $(DT(MLP\_Output) \cap DT(RBF\_Output) \cap DT(SVM\_Output) \cap DT(MLP\_Map) \cap DT(RBF\_Map) \cap DT(SVM\_Map) \cap only\_powerful\_predictors) = \{rt98, rt179\}$.

It suggests that those codons are highly informative in all scenarios. Meanwhile, VL was the single feature in only_powerful_predictors that occurred in the estimated trees for the Output datasets: $(DT(MLP\_Output) \cap DT(RBF\_Output) \cap DT(SVM\_Output) \cap only\_powerful\_predictors) = VL$.

While VL is highly informative in all Output datasets, no evidence was found that this variable interferes, in any scenario, in a negative way, in the process of predicting patient response to antiretroviral therapy, given that it does not appear in either error_predictors or DT(Map) sets.

While ambiguous codon rt210 was selected in two trees, for datasets MLP_Output and SVM_Output (Figure 3AC). In other words, $(DT(MLP\_Output) \cap DT(SVM\_Output) \cap ambiguous\_predictors) = rt210$. The model estimated for RBF_Output did not consider any ambiguous features (Figure 3B), so $(DT(RBF\_OutOut) \cap ambigous\_predictors) = \emptyset$.

In the estimated error mapping models, codons rt63 and rt245 were the only two ambiguous features occurring in the models estimated for MLP_Map and SVM_Map (Figure 3DF), that is, $(DT(MLP\_Map) \cap DT(SVM\_Map) \cap ambigous\_predictors = \{rt63, rt245\}$.

Still analyzing the same error map models, codon rt162 was the only feature in the only_error_predictors present. Interestingly, $(DT(MLP\_Map) \cap error\_predictors) = (DT(SVM\_Map) \cap error\_predictors)$. The same codon is also part of the tree estimated for RBF_Map (Figure 3E). Therefore this is the only codon in only_error_predictors that appears in all three decision tree models estimated for error prediction.

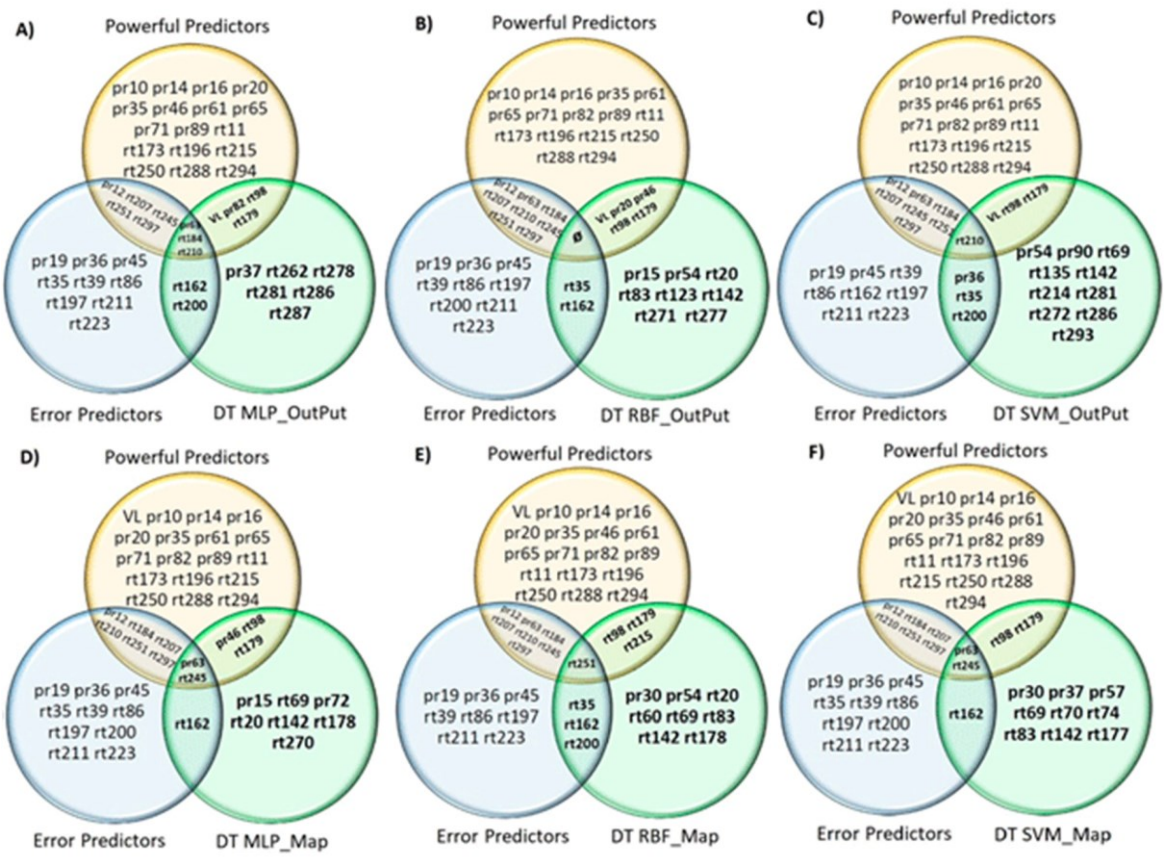

Figure 3. Venn diagram of power predictors, error predictors and features that compose each decision tree model fitted for OutPut and Map datasets.

In general, the trees revealed more features that had not been previously classified than those that had. For example, the tree for MLP_Output (Figure 3A) has six codons, pr37, rt262, rt278, rt281, rt286, and rt87, that do not belong to any of the previous classification lists. The number of attributes not previously classified in each tree is eight (2 PRs + 6 RTs) for RBF_Output, ten (2 PRs + 8 RTs) for SVM_Output, seven (3 PRs + 4 RTs) for MLP_Map, eight (2 PRs + 6 RTs) for RBF_Map, and nine (3 PRs + 6 RTs) for SVM_Map. In all cases, the unclassified features had a lower number of PR codons than that of RT codons.

The Venn diagrams for the features of the trees estimated for Original_Set (A) and EasyHard_Set (B) are shown in Figure 4.

According to the results, $DT(Original\_Set) \cap ambiguous\_predictors = \emptyset$ and $DT(EasyHard\_Set) \cap ambiguous\_predictors = rt245$. Both trees presented the only_error_predictors codons rt35 and rt162, that is, $(DT(Original\_Set) \cap error\_predictors) \cap (DT(EasyHard\_Set) \cap error\_predictors) = \{rt35, rt162\}$.

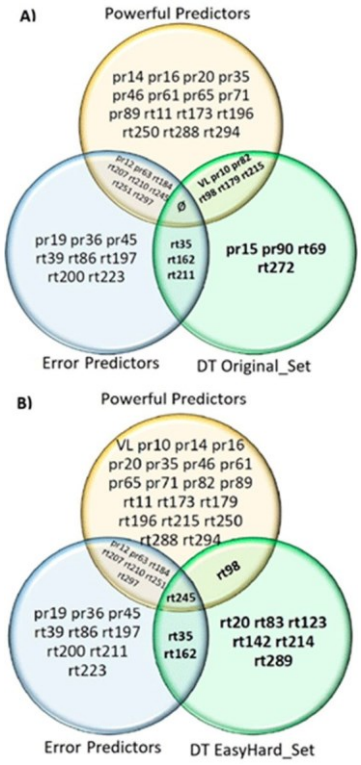

Figure 4. Venn Diagram of Power Predictors, Error Predictors and features that compose each decision tree model fitted for Original_Set and EasyHard_Set.

The variables obtained in the decision model tree fitted for Original_Set represented 2 PRs and 2 RTs (in a total of 4) not previously classified, while the model for EasyHard_Set showed 6 RT codons not previously classified. This result suggests that RT codons are more likely to cause MLM to fail.

Discussion

In our experiments, TCD4+ cell count at the beginning of the treatment was considered by the feature selection method as one of the less informative attributes. When HIV attacks the immune system, profound aggression occurs. At the beginning of the treatment, the VL drops, but the defense system slowly progresses to immune reconstitution. In some cases, especially during the latency period, shortly after the seroconversion, a high number of TCD4+ cells may occur in concomitance with an elevated VL. In these cases, if there is treatment, even decreasing the VL, the amount of this type of cell will not suffer a significant increase.

Table 5. Occurrence of amino acids (AA) in ambiguous codons of PR and RT. Light color marks the presence of an AA in the codon and dark color marks the most frequent AA in that codon.

AA	pr12	pr63	rt184	rt207	rt210	rt245	rt251	rt297
A								
C								
D								
E								
F								
G								
H								
I								
K								
L								
M								
N								
P								
Q								
R								
S								
T								
V								
X								
Total	**14**	**13**	**6**	**11**	**6**	**13**	**9**	**12**

We applied the feature selection technique to identify the most informative variables to be used in the construction of decision tree models. The features that were simultaneously selected for all MLM Output (and original data [Original_Set]) are considered the most informative (powerful_predictors), whereas the variables that were common to all error maps can characterize with precision when determined MLM tends to error (errors_predictors). The features common to these two groups, that is, the intersection between powerful_predictors and errors_predictors, we called ambiguous (ambiguous_predictors). The rt184 was identified as ambiguous, which corroborates our previous results (Rosa et al., 2014), where we concluded that this attribute correlates with the occurrence of MLM errors, but here we found evidence that rt184 is also informative in the prediction of patient's response to treatment. Such codon is widely cited in the literature as related to virus resistance to antiretroviral treatment (Anta et al., 2013; Kulkarni et al., 2012; Ozahata et al., 2015; Xu et al., 2013). In this study, we identified seven other codons that present ambiguous behavior like rt184; they are pr12, pr63, rt207, rt210, rt245, rt251, and rt297. These features showed evidence of being able to predict the patient's response to therapy, but at the same time may lead MLM to failure. We will discuss that further in the journal version of this paper.

We assessed the amino acid composition of the PR and RT codons identified as ambiguous. Table 5 shows the amino acids that occur in each of these eight codons. The lowest amount of amino acid variations occurs in rt184 and rt210 (6), whereas the highest occurs in pr12 (14). In the journal version of this paper, we will discuss the

correlation between the occurrence of MLM errors and the frequency of amino acids in these ambiguous codons.

In our previous study (ROSA et al., 2014), we found evidence that the pr10 codon is highly informative in the prediction process. The results of the present work corroborate this information and also identifies 20 new features of predictive power similar to pr10, which are: VL, pr14, pr16, pr20, pr35, pr46, pr61, pr65, pr71, pr82, pr89, rt11, rt98, rt173, rt179, rt196, rt215, rt250, rt288, and rt294. Also, we identify 11 features that are strongly correlated with the occurrence of MLM errors, namely: pr19, pr36, pr45, rt35, rt39, rt86, rt162, rt197, rt200, rt211 and rt223. In the decision trees construction, we considered first only the only_powerful_predictors, but we observed that the models lost generalization capacity, classifying the instances randomly (data not shown). We then realized that the interaction of the only_powerful_predictors with features not contained in it was more informative than if we considered only the features contained in that set. Thus, we estimated the models without considering the previous classification of features that we performed, but all the features that the Feature Selection method indicated as informative. Analyzing the intersection of the features contained in the trees and those that we previously classified (Figures 3 and 4), we showed that the J48 algorithm considered only a subset of only_powerful_predictors and added to the model codons that had not previously been classified since alone they did not present evidence of information gain. In the next step of this work, we will apply the statistical test χ^2 to verify if there is statistical evidence of interaction between the features: 1) within the same class; 2) between feature classes, and 3) among features without evidence of information gain. These results will be reported in the journal version of this paper.

Conclusion

Finally, we believe that the insights presented in this work can be helpful in the design of new approaches for the prediction of patient's response to antiretroviral treatment and can provide information about the connection genotype-phenotype in HIV.

Acknowledgments

We acknowledge the importance of the data provided by the Resistance Drug Database (https://hivdb.stanford.edu/) to this work. RSR would like to thank Brazilian Sponsoring Agency CNPQ for their financial support.

References

ANTA, L. et al. 2013. Rilpivirine Resistance Mutations in HIV Patients Failing Non-Nucleoside Reverse Transcriptase Inhibitor-Based Therapies. AIDS, v. 27, n. 1, p. 81-85.

AVERT FUNDATION. 2017. Global HIV and AIDS Statistics.

BEERENWINKEL, N. et al. 2005. Computational methods for the design of effective therapies against drug-resistant HIV strains. Bioinformatics, v. 21, n. 21, p. 3943-3950.

CHAWLA, N. V et al. 2002. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Int. Res., v. 16, n. 1, p. 321-357.

DEFO, D.; KOUOTOU, E. A.; RICHIE, J. 2017. Failure to return to receive HIV-test results: the Cameroon experience. BMC Research Notes, v. 10, n. 1, p. 309-314.

EGBE, T. O. et al. 2017. Cesarean delivery technique among HIV positive women with sub-optimal antenatal care uptake at the Douala General Hospital, Cameroon: case series report. BMC Research Notes, v. 10, n. 1, p. 332-34.-

HALL, M. A. et al. 2009. The WEKA data mining software: an update. SIGKDD Explorations, v. 11, n. 1, p. 10-18.

KULKARNI, R. et al. 2012. The HIV-1 Reverse Transcriptase M184I Mutation Enhances the E138K-Associated Resistance to Rilpivirine and Decreases Viral Fitness. JAIDS Journal of Acquired Immune Deficiency Syndromes, v. 59, n. 1, p. 47-54.

KUMAR, V.; MINZ, S. 2014. Feature Selection: A literature Review. Smart Computing Review, v. 4, n. 3, p. 211-229.

LARDER, B. et al. 2007. The development of artificial neural networks to predict virological response to combination HIV therapy. Antiviral Therapy, v. 12, n. 1, p. 15-24.

MAY, M. T. et al. 2014. Impact on life expectancy of HIV-1 positive individuals of CD4 R cell count and viral load response to antiretroviral therapy. AIDS (London, England), v. 28, p. 1193-1202.

OZAHATA, M. C. et al. 2015. Data-intensive analysis of HIV mutations. BMC Bioinformatics, v. 16, n. 1, p. 35-58.

R DEVELOPMENT CORE TEAM. 2018. R: A Language and Environment for Statistical ComputingVienna, Austria. Disponível em: http://www.r-project.org.

REVELL, A. D. et al. 2013. Computational models can predict response to HIV therapy without a genotype and may reduce treatment failure in different resource-limited settings. Journal of Antimicrobial Chemotherapy, v. 68, n. 6, p. 1406-1414

RHEE, S.-Y. et al. 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Research, v. 31, n. 1, p. 298-303.

ROSA, R. S. et al. 2014. Insights on prediction of patients' response to anti-HIV therapies through machine learning. 2014 International Joint Conference on Neural Networks (IJCNN). Anais...IEEE, jul. 2014. Disponível em: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6889659.

WANG, D. et al. 2009. A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy. Artificial Intelligence in Medicine, v. 47, n. 1, p. 63-74.

XU, H.-T. et al. 2013. Effect of mutations at position E138 in HIV-1 reverse transcriptase and their interactions with the M184I mutation on defining patterns of resistance to nonnucleoside reverse transcriptase inhibitors rilpivirine and etravirine. Antimicrobial agents and chemotherapy, v. 57, n. 7, p. 3100-9.