Final Project Report
Prediction of HIV Progression

Ou Bai, Weiqiang Liu, Yuzhong Qian and Fanjie Xiao

Abstract

Inspired by the condition that there are many people suffering the HIV, we developed a way that can identifies the accurate rate of prediction of HIV progression by applying many different machine learning. By comparing the result, we can know that applying Naive Bayes Classifier can achieves the best performance. We also developed a method that can identify the markers that can lead to the severity of the infection by utilizing the weights of the features used in the Random Forest algorithm. We performed an evaluation by applying the marker to new DNA sequences and got good results.

Introduction

The human immunodeficiency virus (HIV) is a lentivirus (a subgroup of retrovirus) that causes HIV infection and over time acquired immunodeficiency syndrome (AIDS). AIDS is a condition in humans in which progressive failure of the immune system allows life-threatening opportunistic infections and cancers to thrive.[1] According to the World Health Organization(WHO), HIV has caused 25 million deaths worldwide since it was first recognized in 1981. Moreover, the virus will likely evolve around recent drugs, making it crucially important to get a better understanding of the virus.[2] We know that without treatment, HIV advances in stages, overwhelming your immune system and getting worse over time. The three stages of HIV infection are: (1) acute HIV infection, (2) clinical latency, and (3) AIDS. However, there's good news: by using HIV medicines consistently, you can prevent HIV from progressing to AIDS. However, not everyone is diagnosed early. Some people are diagnosed with HIV and AIDS concurrently, meaning that they have been living with HIV for a long time and the virus has already done damage to their body by the time they find out they are infected. [3] If we can predict the stages in advance by analyzing the DNA sequences, you will most likely never progress to AIDS. What's more, if we can find the the markers that may lead to HIV's getting worse, we can know the tendency of patient's condition and we can use different drugs accordingly.

In this project, we are provided with the data that is comprised of 1000 records of patients. The data provides the nucleotide sequences of their Reverse Transcriptase(RT), their Protease(PR) and their viral loads and CD4 count at the beginning of therapy and the result whether HIV is getting better or worse. The PR sequence is the blueprint of the protein, which is also the workhorse of the cell. The RT sequence is also doing the task of copying the HIV-1 genome within the cell. The viral load is the term to describe the amount of HIV in a body fluid. The CD4 count is a test that measures how many CD4 cells you have in your blood. We use 1 to represent that the HIV is getting better and 0 to represent that the HIV is getting worse. We will split the data into training groups and testing groups. All the data comes from Kaggle. [4]

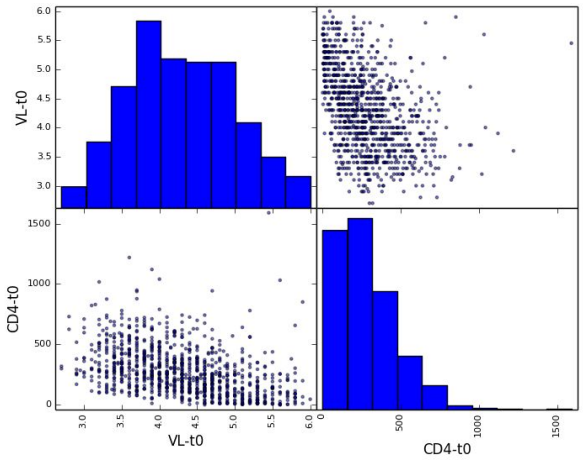


Figure 1: The distribution of CD4 and VL-t0.

Methods

1. Train the models and predict

We split the data into training data (80%) and testing data (20%). In order to get rid of the overfitting problem, we have utilized n-fold cross validation (n = 10) on the training data. First we ignore those does not have the PR sequence or the RT sequence. Since the lengths of each PR sequence and RT sequence are different, we have used a multiple sequence alignment vis MUSCLE[5] to do this job, so that we can get the sequences with the same length. As for the models, we have built three models which are comprised of three different feature groups. The first one contains the following two features: the the PR sequence and RT sequence from the virus. The second one contains other two features: the VL and CD4 from the patients. The third one is the combination of the above features. As to the algorithms, we have applied Logistic Regression(LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Decision Trees, Gaussian Naive Bayes Classifier (NB) and Support Vector Machine (SVM), Random Forest (RF) on the dataset to train the three different models. The evaluation method we used is the accurate rate, which calculates the number of correct predictions as a proportion of the total number of the predictions.

2. Finding the markers

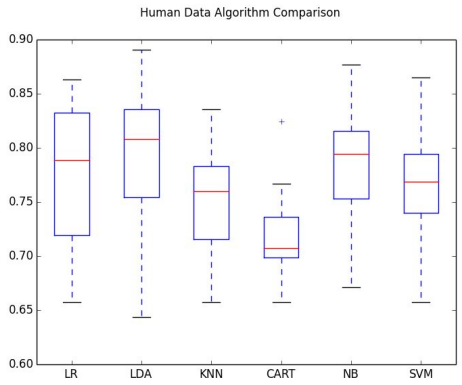
We used the byproduct of machine learning to solve this problem. Firstly, We used the python library to convert the string of PR and RT sequences to vectors via One Hot Representation. And we used all the features including the viral load and CD4 count to train the models by applying the Random Forest algorithm. After training the models, we can get the weight for each feature. By specifying the length of the markers, we can get the weight for all the possible markers of the fixed length. After sorting them, we can know which position plays the most important role in the DNA sequence. And then we traversed all the sequences, and defined the positive markers as the subsequences in the important positions such that the positive records with the subsequence are much more than the negative records with it. With the similar method, we got the positive and negative markers of PR and RT sequences.

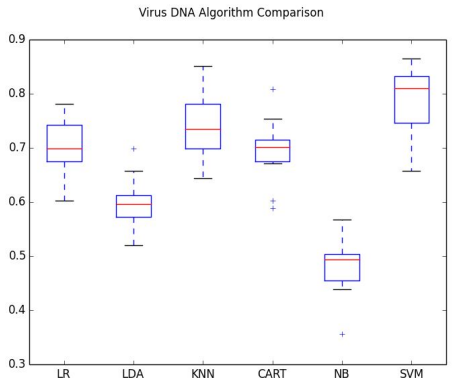
Results & Discussion

1. Based on the results of three models, we find that the one that takes four attributes into consideration can obtain the highest accuracy. Therefore, we will use all of our attributes to find markers. Below is the result table:

	Virus DNA Sequences		Human Body Information		All Attributes	
	Cross Validation Accuracy	Testing Data Accuracy	Cross Validation Accuracy	Testing Data Accuracy	Cross Validation Accuracy	Testing Data Accuracy
LR	0.702073	0.677596	0.777194	0.825137	0.801925	0.836066
LDA	0.597131	0.551913	0.789522	0.792350	0.635265	0.612022
KNN	0.740244	0.803279	0.752610	0.775956	0.751240	0.775956
CART	0.707571	0.715847	0.717049	0.704918	0.765031	0.704918
NB	0.480674	0.437158	0.777231	0.754098	0.773158	0.846995
SVM	0.788116	0.836066	0.770363	0.770492	0.788116	0.836066
RF	0.748445	0.721331	0.751351	0.792350	0.770344	0.830601

In the table, Gaussian Naive Bayes Classifier obtains the best result (shown in the red figure) when it takes all the attributes as input. So we will use NB as our prediction function.





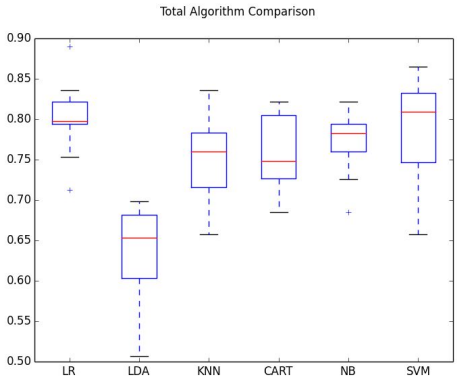


Figure 2: The performance of different algorithms.

Figure 2 are the details of Cross Validation. We separated training data into 10 folds and got the results range and average value.

2. After using the byproduct of machine learning, we can find many markers and their positions. Figure 3 is the result for all the possible markers with large possibility. Then we use these possible markers to predict the sequences again and we achieve the 77% accuracy.

```
PR_pos
[['CTYGTCACAATAAAG', 27, 41], ['CTCGTCACAATAAGG', 27, 41], ['CTCGTCACAATAAAR', 27, 41],
['AGRTAGGGGGGCAAC', 40, 54], ['AGATAGGGGGGCARC', 40, 54], ['AGATAGGGGGGCAAA', 40, 54],
['AGATAGGGGGGCAAT', 40, 54], ['ARATAGGGGGGCAAC', 40, 54], ['YGTCACAATAAAGAT', 29, 43],
['CGTCACAATAAGGAT', 29, 43], ['CGTCACAATAAAGRT', 29, 43], ['CGTCACAATAAARAT', 29, 43], ...]
PR_neg
[['GAGGTGGGGATTGAC', 629, 643], ['GAGGTGGGGACTTAC', 629, 643], ['AAGTGGGGGGTTTACC', 630, 644],
['AGGTGGGGACTTACC', 630, 644], ['AGGTGGGGATTGACC', 630, 644], ['GGTGGGGATTGACCA', 631, 645],
['AGTGGGGGGTTTACCA', 631, 645], ['GGTGGGGACTTACCA', 631, 645], ['TGAAGTGGGGGGTTTT', 628, 642],
['TGAGGTGGGGATTGA', 628, 642], ['TGAGGTGGGGACTTA', 628, 642], ['AGTACCACTAACAAG', 875, 889], ...]
RT_pos
[['CTCGTCTCAATAAAG', 27, 41], ['ATCGTCACAATAAAG', 27, 41], ['ATCGTCACAGTAAAG', 27, 41],
['CTCGTCACAATAAAA', 27, 41], ['AGATAGGGGGGCAAY', 40, 54], ['AAATAGGGGGGCAAC', 40, 54],
['AGATAGGAGGGCAAC', 40, 54], ['AGGTAGGGGGGCAAC', 40, 54], ['CGTCACARTAAAGAT', 29, 43],
['CGTCACAATAAAAAT', 29, 43], ['CGTCTCAATAAAGAT', 29, 43], ['CGTCACAGTAAAGGT', 29, 43], ...]
RT_neg
[['GAGGTGGGGATTAAC', 629, 643], ['AAGGTGGGGACTTAC', 629, 643], ['GAGGTGGGGATTCTT', 629, 643],
['AAGGTGGGGATTTAC', 629, 643], ['GAAGTGGGGATTCAC', 629, 643], ['GAAATGGGGGTTTTA', 629, 643],
['AAGGTGGGGATTTTA', 629, 643], ['GAGGTGGGGGTTTTA', 629, 643], ['AGGTGGGGGTTTTAC', 630, 644],
['AAATGGGGGTTTTAC', 630, 644], ['AGGTGGGGACTTTTC', 630, 644], ['AGGTGGGGATTCTTC', 630, 644], ...]
```

Figure 3: The possible markers and their positions.

3. After getting the markers, we can predict the progression of the patients. According the markers, we can treat the patients with the specific effective drugs.

Conclusions

The prediction of HIV progress is beneficial to human beings. By applying machine learning techniques, we can predict it with high accuracy. And based on the byproduct of machine learning, we predicted the markers that are useful for predicting the progress. Once we know the

sequence of markers, we can get protein structure from these markers. Drug companies can utilize these protein structure to produce drug for patients in different stages.

References

[1] HIV, https://en.wikipedia.org/wiki/HIV
[2] Kaggle Competition, https://www.kaggle.com/c/hivprogression#description
[3] Stages of HIV Infection, https://www.aids.gov/hiv-aids-basics/just-diagnosed-with-hiv-aids/hiv-in-your-body/stages-of-hiv/
[4] Kaggle Data, https://www.kaggle.com/c/hivprogression/data
[5] Muscle Alignment, http://drive5.com/muscle/

Role of each student

Ou Bai and Yuzhong are responsible for training the models and getting the accurate rate. Fanjie, Weiqiang and Yuzhong are responsible for finding markers in the HIV DNA sequences.