Project Update

Ou Bai, Weiqiang Liu, Yuzhong Qian and Fanjie Xiao

Aim/Hypothesis

1. In this project, we aim to analyze the progression of the HIV and predict the likelihood that an HIV patient's infection will become less severe in order to help patient better. [1]
2. What's more, we are aiming to find the markers in the HIV sequences that will mostly lead to the severity of infection so that we can know the tendency of patient's condition and we can use different drugs accordingly.

Computational Approaches developed

1. Building models and predict
We split the data into one group of training data (80%) and one group of 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. Since the length of each PR sequence and RT sequence is different, we have used one multiple sequence alignment called muscle[2] 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 is comprised of three different feature groups. The first one is the the PR sequence and RT sequence from the virus. The second one is 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) on the dataset to train the three different models. The evaluation method we have developed is the accurate rate, which calculates the number of correct predictions as a proportion of the total number of the predictions.

2. Finding markers in the HIV sequences
Since we have already finished up the multiple sequence alignment in the above section, the next step we have implemented is to cluster the sequences into different clusters that can represent for different stages based on the the result of the HIV progression. We will cluster the part of data where the HIV conditions of these patients are getting better and cluster other data when the HIV conditions of patients are turning worse. We have used the position-specific scoring matrix (PSSM) to get the sequence with the largest possibility for each cluster, which can represent for all the sequences in that cluster. We try to compare these representatives and find the differences between them. Then we got the unique marker that is in some specific positions which can be used to know the tendency of the patients' conditions and provided them to the pharmaceutical factories for better treatments to patients.

Data Used

The data is comprised of the records of 1,000 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. 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. [3]

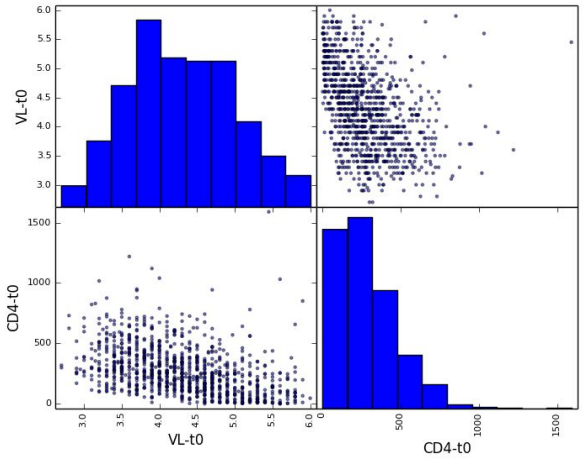

Figure 1: The distribution of CD4 and VL-t0.

Current Results

1. Based on the results of three models, we find that the one taking 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

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

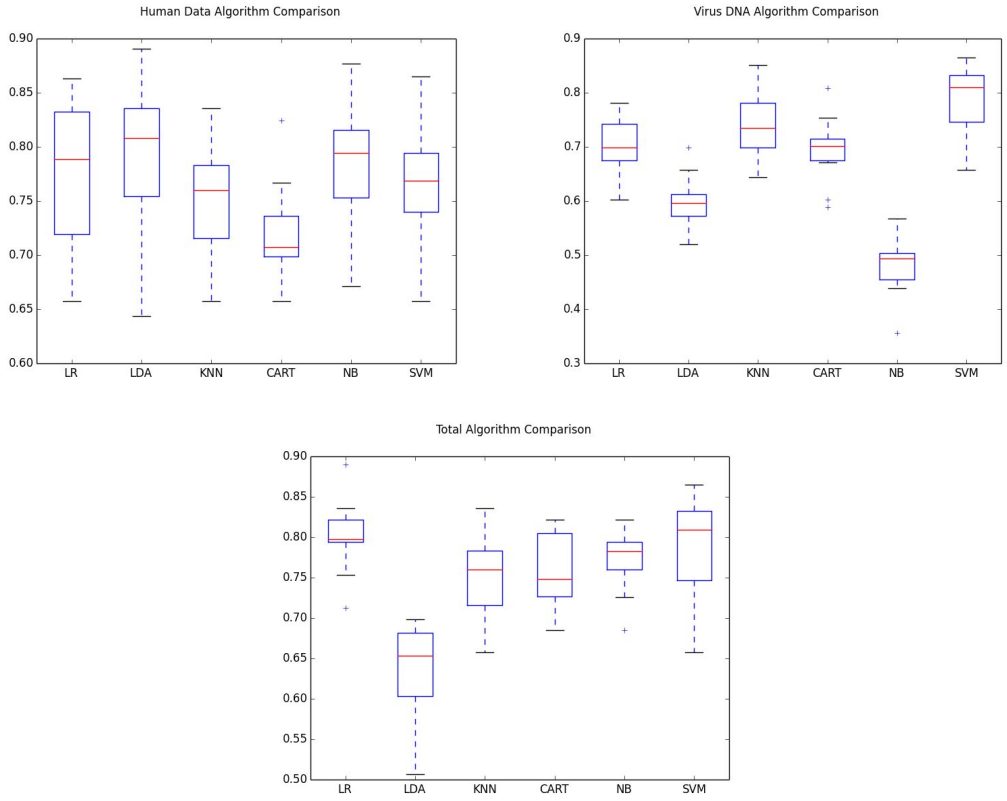

Figure 2: The performance of different algorithms.

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

2. We have used the CD-HIT, which stands for Cluster Database at High Identity with Tolerance, developed by the Burnham Institute (now Sanford-Burnham Medical Research Institute)[4], to cluster the data and got one representative for each cluster. Figure 3 is calculated when the sequence identity parameter is 90%. Figure 3(a) is showing part of the clusters, which include the number of sequences in the clusters and their representatives. Figure 3(b) is the distribution of the clusters.

Cluster 0, No. sequences: 133, Representative: 610, length:297
Cluster 1, No. sequences: 121, Representative: 822, length:297
Cluster 2, No. sequences: 118, Representative: 53, length:297
Cluster 3, No. sequences: 87, Representative: 267, length:297
Cluster 4, No. sequences: 47, Representative: 861, length:297
Cluster 5, No. sequences: 46, Representative: 529, length:297
Cluster 6, No. sequences: 37, Representative: 886, length:297
Cluster 7, No. sequences: 36, Representative: 89, length:297
Cluster 8, No. sequences: 23, Representative: 732, length:297
Cluster 9, No. sequences: 22, Representative: 259, length:297
Cluster 10, No. sequences: 19, Representative: 26, length:297
Cluster 11, No. sequences: 17, Representative: 728, length:297
Cluster 12, No. sequences: 17, Representative: 106, length:297
Cluster 13, No. sequences: 16, Representative: 246, length:297
Cluster 14, No. sequences: 16, Representative: 134, length:297
Cluster 15, No. sequences: 16, Representative: 69, length:297
Cluster 16, No. sequences: 15, Representative: 685, length:297
Cluster 17, No. sequences: 14, Representative: 202, length:297
Cluster 18, No. sequences: 12, Representative: 110, length:297
Cluster 19, No. sequences: 10, Representative: 256, length:297
Cluster 20, No. sequences: 9, Representative: 157, length:297
Cluster 21, No. sequences: 8, Representative: 237, length:297
Cluster 22, No. sequences: 8, Representative: 715, length:297
Cluster 23, No. sequences: 8, Representative: 823, length:297
Cluster 24, No. sequences: 7, Representative: 28, length:297
Cluster 25, No. sequences: 7, Representative: 130, length:297
Cluster 26, No. sequences: 7, Representative: 731, length:297
Cluster 27, No. sequences: 6, Representative: 918, length:297
Cluster 28, No. sequences: 5, Representative: 228, length:297
Cluster 29, No. sequences: 5, Representative: 664, length:297

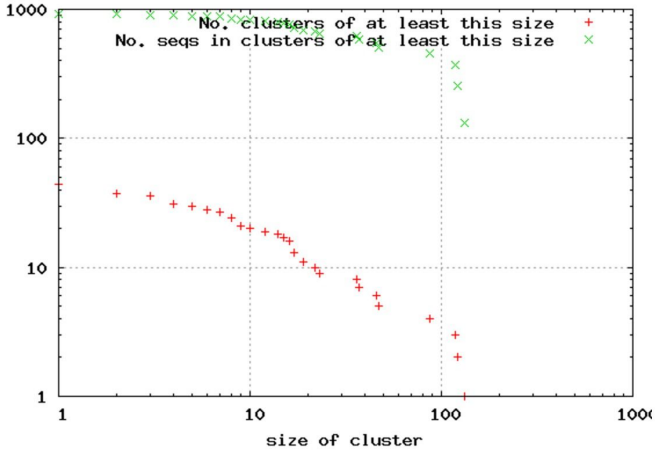

Figure 3: The information for each cluster

To Do List

1. Currently, we cluster the data into several groups. Future work is we will use these groups to find markers. And the way to find markers we try to implement is multiple sequence alignment plus PSSM.

2. Using our prediction model and our markers to predict the progression of new HIV DNA sequence and compare these two results to verify whether we find the effective markers or not.

Expected Results

We use our model to predict amount of new coming sequences and get the result of getting better or worse. And try to find the relative markers or motifs in each sequence. If many of them match the markers we found in the training data, it means the markers are correct and we can predict the HIV progression by identifying whether the DNA has these markers. And these markers also can be used in the drug production to help patients.

References

[1] Kaggle, https://www.kaggle.com/c/hivprogression#description
[2] Muscle, http://drive5.com/muscle/
[3] Data, https://www.kaggle.com/c/hivprogression/data
[4] CD-HIT, http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi