A PROJECT REPORT ON USING HIDDEN MARKOV MODEL TO CLASSIFY VIRUSES

Presented to

Dr. Leonard Wesley

Department of Computer Science

San Jose State University

In Partial Fulfillment

Of the Requirements for the Class

CS 223

By

Divya Singh (010728445)

Krishna Teja Vojjila (011410516)

Tharun Theja Kammara (010953280)

**TABLE OF CONTENTS**

I. Introduction …....…....………………………………………………………………………...2

II. Background ....………………………………………………………………………………..3

III. Application and Design ........……………………………………………………………….4

1. Dataset …....……………………………………………………………………….........4
2. Workflow ……………....………………………………………………………………...5
3. Build and train the HMM ………………………………………………………............6
4. Test the HMM …………………………………………………………………………...7

IV. Example and Results ……………………………………………………………………….7

V. Future Work………………………………………………………………………………….11

VI. References …………………....…………………………………………………………...12

VII. Member Contribution ………....………………………………………………………….13

VIII. Appendix: How to run the program …....………………………………………………14

**Introduction**

Viruses are small infectious agents that were not witnessed until 1931. They are one-hundredth size of the bacteria and requires an electron microscope to see them. They survive by infecting living host cells and replicating their DNA/RNA via their host. The structure of viruses vary widely between different families and strands. Viral infections are a major global health concern, and new infectious diseases continue to emerge [1]. These diseases are a tremendous burden on economies and public health, so there is a high demand for advances in viral diagnostic methods. Detection of viruses in specimens traditionally depends on amplification of conserved regions of nucleic acid from viral genomes. These traditional approaches have limitations.

Recent research work on Hidden Markov Model (HMM) analysis has shown that it is an effective tool in detecting viruses and classifying them. The goal of this project is to classify genomic models as virus structures using a viral and non viral genomic dataset. The main purpose is to get a sight into different methods common in the bioinformatic field and find optimal solutions to existing problems. By training and fitting the model with viral genomic dataset, the nucleotide composition of viral genomes is known and the genomic parameters are estimated that are to to be considered for testing our unknown genomic sequence. HMM is trained using viruses and the trained models are able to classify a particular virus family from non-viral programs. The three main questions on HMM in our model i.e. evaluation, decoding and learning are also addressed.

The report is organized as follows. Background discusses the previous

Approaches used in classifying viruses. Application and design describes the dataset used in training the HMM model and the architecture of HMM along with the steps that are followed in building and testing the model. Example and Results gives the snapshots of results obtained while training and testing the model. It is followed by Future Work, References and Member Contribution. Then, Appendix guides how to run the program.

**Background**

A number of successful attempts have been conducted where machine learning methods are used to identify, classify and predict viruses.

In [2], to improve influenza surveillance and vaccine development, a series of bioinformatic analyses is applied to characterize currently circulating influenza viral strains and identify their evolutionary origin. Machine Learning techniques such as Support Vector Machine (SVM), Decision tree and phylogenetic analysis are used to classify viruses among human, human wine and latest pandemic human/swine. Based upon the results obtained in analyses, a web tool is created that uses HMM to predict influenza H1V1 accurately.

Human Immunodeficiency Virus (HIV) causes AIDS. HIV is of two types-HIV-1 & HIV-2. HIV is different in structure from other retroviruses. [3] developed a machine learning model to classify alpha, beta, and residues of four types of HIV enzymes that are present in HIV1 and HIV2 cycle. Random Forest, Rotation Forest and J48 are used to classify these residues and the model developed gives fair accuracy [3].

[4] classified dengue fever patients based on gene expression data using SVM. It analyzed the expression pattern of 12 genes in peripheral blood mononuclear cells (PBMCs) of 28 dengue patients (13 Dengue haemorrhagic fever and 15 Dengue fever) during acute viral infection. The model is trained using gene expression data of these genes and achieved the highest accuracy of 85% with leave-one-out cross-validation.

In [5], Alternating Decision Trees (ADT) are used to identify family of viruses. It used Adaboost, a supervised machine learning algorithm, to train an ADT on a given

Dataset. ADT minimizes the exponential loss using coordinate descent algorithm.

**Application and Design**

HMM method has been traditionally used in signal processing, speech recognition, and, more recently, bioinformatics. In bioinformatics, it is used in [sequence alignment](http://www.bioinformatics.org/wiki/Sequence_alignment), [gene detection](http://www.bioinformatics.org/w/index.php?title=In_silico_gene_detection&action=edit&redlink=1), [structure prediction](http://www.bioinformatics.org/wiki/Structure_prediction), [data-mining literature](http://www.bioinformatics.org/w/index.php?title=Data-mining_literature&action=edit&redlink=1), and so on. Difficulties with this method include the need for accurate, and sufficiently sized [training](http://www.bioinformatics.org/w/index.php?title=Training_set&action=edit&redlink=1) dataset. In case of gene detection, in order to accurately predict genes in the [human genome](http://www.bioinformatics.org/w/index.php?title=Human_genome&action=edit&redlink=1), many genes in the genome must be accurately known. Similarly, it is necessary for the model to understand how a virus looks like i.e. parameters (transition and emission probabilities) in order to accurately classify a sequence as virus or non virus.

**1. Dataset**

The dataset is obtained from GenBank Database. The link for dataset is provided as follows: <https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239>. There are several species among which viruses can be found but for now we are only considering Banana streak viruses in plants. Banana streak viruses are most common type of pathogens in bananas. There are 10 families of banana streak viruses. The training data contains 7500 nucleotides each. The data is cleaned by removing unwanted parameters, spaces and punctuations.

The example of one of the dataset is given below:

*1 tggtatcaga gcaaggttag ttcttatggc tttcatgggg taaaaccctt aggtaggagc*

*61 cgatgggctc tgctattttt gatttgggtt aatggttgta caagttctat gatagataag*

*121 tcgaatggag caacacttat gtatgaaaaa tgattgccta tgataataga tgggctaagg*

where numbers in the beginning of each genome sequence represent nucleotide index.

**2. Workflow**

The following steps are taken in design of the process to be used for training and testing the HMM model:

1. Consider models applying to gene annotation in viruses.
2. Train the model.
3. Calculate and select genomic parameters from the trained model.
4. Create model containing those parameters.
5. Fit the above model to GenBank.
6. Estimate parameters and evaluate results.

**3. Build and train the HMM**

A Hidden Markov Model is statistical model that has hidden states and known probabilities of the state transitions.

According to definition, HMM is a 5-tuple where,

1. Q is a finite set of states
2. A finite set of alphabets
3. State Transition Function
4. State Emission Function
5. q0, Start State

Our model architecture is shown in Fig.1.

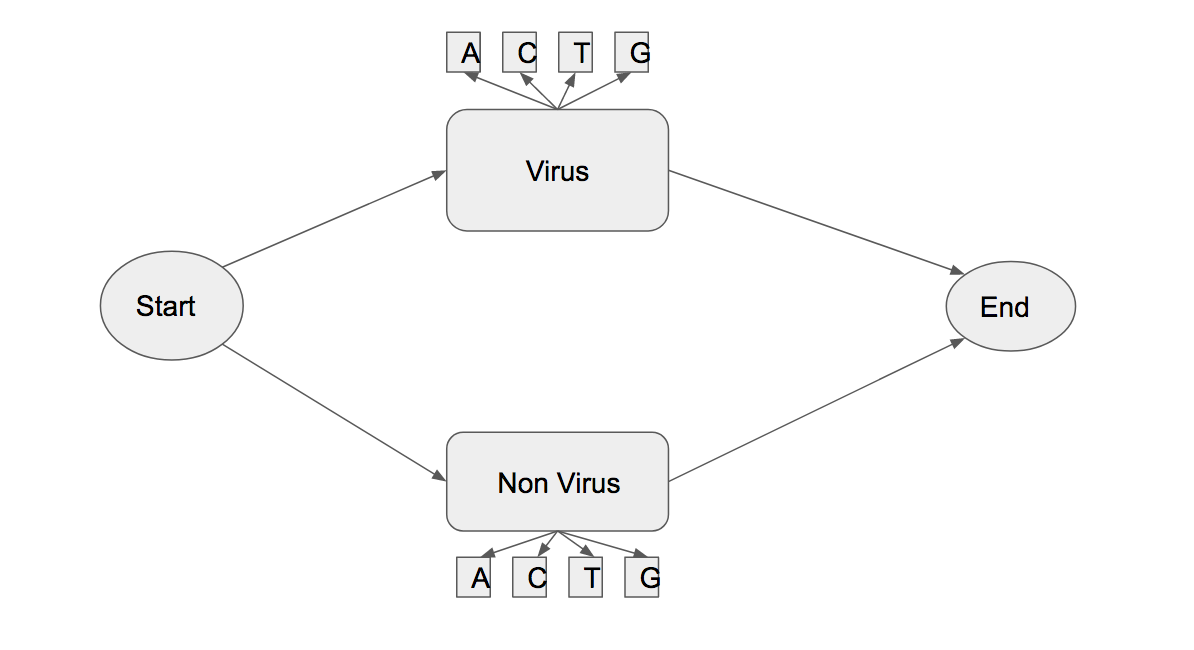


Fig.1. Hidden Markov Model containing four states

A HMM containing 4 states i.e. Start, Virus, Non Virus, and Exit. The model is trained on training dataset to estimate the best parameters like start, emission and transition probabilities. This is the learning problem of HMM i.e. The Baum-Welch Algorithm. Then we use Viterbi algorithm to determine the path of hidden states that classify the genome as virus. This is the decoding problem of HMM.

**4. Test the HMM**

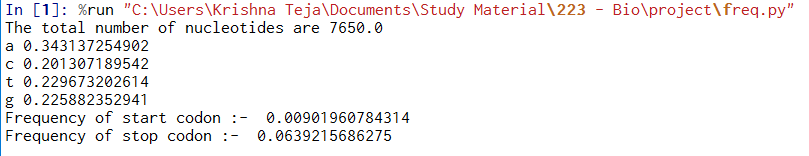
After getting a model, further we test the models on our test dataset. The model is tested on two genome sequences. One sequence is taken from dataset Banana streak virus and another sequence is taken from non virus family (in this case bacteria is used). The test dataset is also obtained from the GenBank database. The dataset contains about 7801 nucleotides. The model gives higher probability for sequence that

belongs to banana streak virus thus classifying it as a banana streak virus. This is the evaluation problem of HMM i.e. Forward Algorithm.

**Example and Results**

**Calculating ACTG frequencies:**

For the initial emission matrix, ACTG frequencies are taken in the dataset by taking the genome as a string and iterating over it and counting the respective counts of A,C,T,G over the total number of nucleotides.

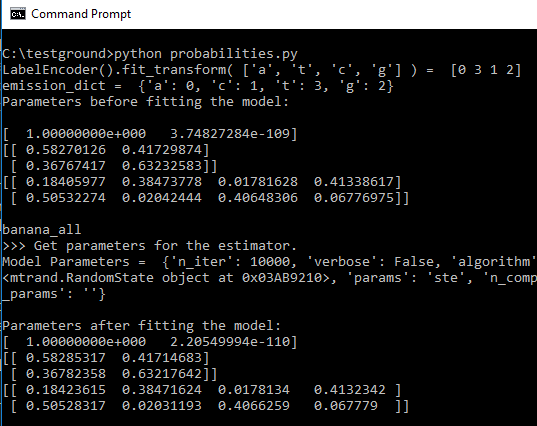
For the parameters, the frequency of start and stop codons is considered which are calculated over the training data as given in the below snapshot.

**Calculating transition and emission probability matrices:**

Initially, a simple hmm script is created and the mutlinomial\_hmm method of *hmmlearn* is run using params=”ste” and that will generate the emission and transition probabilities matrices from the observed sequences.



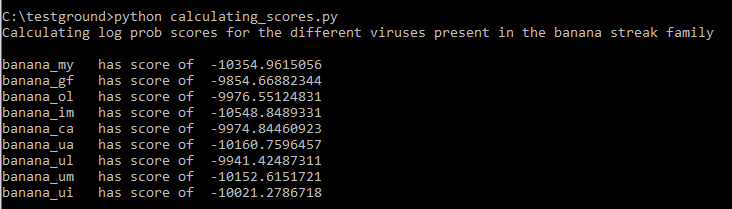
The program is iterated 10000 times for increased accuracy of the matrices. All the genomes are added from the 10 banana streak virus families and are fed as input to hmm model. The emission and transition probabilities are obtained as given in the following snapshot.



Now, the parameters obtained are used in hmm model for training and testing.

**Training:**

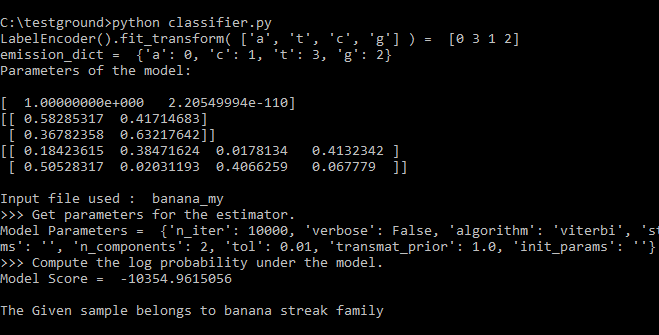
The calculating\_scores.py file is used to train the hmm model. The hmm

model is built with the above obtained matrices and log prob scores are calculated for all the different sub viruses in the banana streak family as shown in the snapshot given below. 

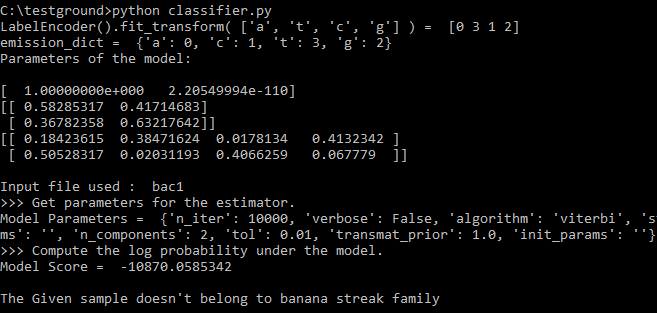
**Testing :**

Now, the log prob scores obtained for different viruses of banana streak family in the above training phase are used to determine whether a given sample belongs to banana streak family or not.

*a) Using a genome sequence from the banana streak virus family for testing*



*b) Using a genome sequence from other than the banana streak virus family for testing*



**Conclusion and Future Work**

Thus, in this project HMM model is built to do binary classification that identifies whether a given strand of DNA belongs to the Banana Strain Virus family or to the Aceto Bacteria family based on the ACTG frequencies and the Intron Exon frequency taken as parameters. The log score is used as an evaluation matrix.

As the broad topic of classification is dealt, there is a lot of scope in future where several additional parameters can be considered . for classification such as length of introns and exons, etc. And also by training model on a larger training dataset and working on tuning the model can help in improving the accuracy score.

Further, the project can be extended by training several HMM models on different families of viruses and hence classifying a lot more DNA sequences. This surely has a lot of potential to solve many problems in the fields of healthcare, medicine research, etc.

**References**

[1] Dong, J., Olano, J.P., McBride, J.W. and Walker, D.H., 2008. Emerging pathogens:

challenges and successes of molecular diagnostics. *The Journal of Molecular*

*Diagnostics*, *10*(3), pp.185-197.

[2] Attaluri, P.K., Zheng, X., Chen, Z. and Lu, G., 2009. Applying machine learning

techniques to classify H1N1 viral strains occurring in 2009 flu pandemic. *BIOT-2009*,

*21*.

[3] Dubey, A., Pant, B. and Chouhan, U., 2011. Machine learning model for HIV1 and

HIV2 enzyme secondary structure classification. *J. Comput. Method. Mol. Design*,

*1*(2), pp.1-8.

[4] Gomes, A.L.V., Wee, L.J., Khan, A.M., Gil, L.H., Marques Jr, E.T., Calzavara-Silva,

C.E. and Tan, T.W., 2010. Classification of dengue fever patients based on gene

expression data using support vector machines. *PloS one*, *5*(6), p.e11267.

[5] Raj, A., Dewar, M., Palacios, G., Rabadan, R. and Wiggins, C.H., 2011. Identifying

hosts of families of viruses: a machine learning approach. *PloS one*, *6*(12),

p.e27631.

**Member Contribution**

**Divya Singh:**

Research and Literature review

Chose the dataset

Trained the HMM and extracted the parameters

Report (Introduction, Background, Application and Design, Results)

**Krishna Teja Vojjila:**

Designed the HMM model

Trained the model and extracted the parameters

Calculated ACTG frequencies and frequency of start and stop codons in the dataset

Report (Introduction, Application and Design, Example and Results, Conclusion & Future work)

**Tharun Theja Kammara:**

Designed and applied Baum Welsh to calculate model parameters

Applied Viterbi Algorithm to find out the hidden states

Tested on the sample sequences

Report (Introduction, Application and Design, Example and Results, Future Work)

**Divya, Krishna, Tharun:** Proposal writing, Milestone report writing, PowerPoint Presentation and Report writing

**Appendix: How to run the program**

1) The folder **code\_dir** contains the files with genome sequences of the banana streak family virus (to be used for training and testing).

2) The **banana\_all** contains the appended genome sequences from all the banana streak family viruses used for getting the transition and emission probabilities.

3) The **bac1** is a bacterial genome of Acetobacter aceti used for testing the model.

##python code:

4) **frequency.py**: to calculate the initial frequencies of ACTG and start/stop codons in the banana streak families.

5) **probabilities.py**: python code to find out emission and transition probability matrices.

6) **calculating\_scores.py**: python code to find out the log prob scores of genomes of banana streak family virus.

7) **classifier.py**: python code to classify the given genome sample