Classification of HIV type One Strains Using Profile Hidden Markov Models

A Thesis submitted in partial fulfillment of the requirements for the award of the degree of

Master of Technology

In

Computational and System Biology

Submitted By:

Sanjiv Kumar Dwivedi

E.NO: 09/75/MT/14

Under the Guidance of

Dr. Supratim Sengupta

Associate Professor

Centre for Computational Biology and Bioinformatics

School of Computational and Integrative Sciences

Jawaharlal Nehru University

New Delhi-110067

जवाहरलाल नेहरू विश्वविद्यालय

JAWAHARLAL NEHRU UNIVERSITY

संगणकीय एवं समेकित विज्ञान संस्थान

School of Computational and Integrative Sciences

नई दिल्ली- 110067

NEW DELHI- 110 067 (INDIA)

Hall No-6, Lecture Hall Complex, JNU
Tel. : (Direct) 26741517 (Off.) : 26704171
Fax : 011-26741586
Email : dean_sit@mail.jnu.ac.in

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled "***Classification of HIV Type One Strains Using Profile Hidden Markov Models***" is an authentic record of the dissertation carried out by **Sanjiv Kumar Dwivedi** at School of Computational and Integrative Sciences, Jawaharlal Nehru University, under my supervision and guidance. The contents of this project work, in full or in parts, have not been submitted to any other institute or University for the award of any degree or diploma. The work fulfills the requirement for the award of the M. Tech degree in Computational and Systems Biology.

.....

Dr. Supratim Sengupta
Associate Professor
Supervisor
SCIS, JNU.

.....

Professor Indira Ghosh
Dean
SCIS, JNU.

DEAN
School of Computational & Integrative Sciences
Jawaharlal Nehru University
New Delhi - 110067

ACKNOWLEDGEMENT

With an overwhelming sense of pride and genuine obligation, I express my deepest regards to my supervisor Dr. Supratim Sengupta for providing excellent guidance in proper direction, motivation and encouragement throughout the project work which enabled me to develop a good understanding of the subject. He introduced me in this area of work and has always been with me in any case of need from initial to final stage of the work. Without his constant support and motivation, this research work could not be accomplished successfully.

I gratefully thank to the Dean of our School, Prof. Indira Ghosh for her motivating advice during my course and providing us the best facilities.

I express my sincere gratitude to my teachers Prof. Indira Ghosh, Prof. R. Ramaswamy, Dr. N . Subbarao, Prof. Rahul Roy, Dr. Andrew M Lynn, Dr. A. Krishnamachari, Dr. Lovekesh Vig, Dr. Pradipta Bandyopadhyay, Dr. N.S. Sahni, Dr. Rashi Gupta, Dr. Devapriya Choudhury for providing valuable teaching throughout the course.

I would like to extend the gratitude to my friends and research scholars (Payal, Ashutosh and Abhay) who spend their valuable time for supporting me to finish this work.

I have no words to express my sense of gratitude to my parents for being a constant source of true affection and encouragement without which I could not think to reach this level in my life.

I acknowledge the financial support and resources provided by D.B.T. , U.G.C. and JNU.

Finally, I would like to thank everyone who contributed this successful completion of the work.

CONTENTS

Abstract

The number of sequences of HIV type one viruses are increasing rapidly over time. To accurately classify the newly sequenced genomes into appropriate subtypes is important from the clinical viewpoint. It is essential for understanding how, when and where new variants of HIV-1 arise and for understanding the spatial distribution of each strain. Such information can lead to the development of more focused treatments for patients infected with a specific strain of the virus. In this thesis we show that a method based on profile Hidden Markov Models (pHMM) can accurately classify not only pure strains of HIV1 but also determine the subtype composition of Circulating Recombinant Forms (CRFs) which are made of two or more subtypes.

Chapter 1

Introduction

1.1 Background:

There are more than 31.1 million people living with human immunodeficiency virus type one (HIV-1) and more than 25 million deaths have been caused by the virus all over the world [1]. Infection from HIV is distributed over different zones of the world. The vaccine development strategy faces a major challenge because of the genetic variability of HIV that arises due to the lack of proofreading capability of the reverse transcriptase enzyme [2,3,4].

The database of viral sequences of HIV-1 is increasing over time and therefore an accurate and reliable classification of different strains of HIV-1 is important for many aspects of complex biology of the viruses[5]. Consequently classification of these virus strains into different subtypes based on their genetic dissimilarity plays an important role in understanding their evolution, distribution and geographical spread. It is also crucial for monitoring information about disease transmission by AIDS, to help in developing antiviral therapies and/or vaccines and to take decisions on the treatment strategy[6].

1.2 Basic biology of HIV-1

HIV is a member of the genus Lentivirus , subfamily Lentivirinae and family is Retroviridae.

The internal structural and its components are shown in Figure 1.1. The shape and size of HIV virus is roughly spherical and its diameter is about one tenth of a micrometer. The outermost envelope is composed of a bilayer membrane of lipids that contains numerous spikes. The spikes are embedded in the membrane and are made up of four molecules of glycoprotein gp120 and same number of glycoprotein gp41. A layer of matrix protein surrounds the core (capsid) which in turn surrounded by an envelope. The genetic material of the HIV virus is contained in the hollow truncated cone shape capsid that is composed of another protein, p24. Inside the viral core there exists two strands of RNA contains about 9200 nucleotide bases, a protease, reverse transcriptase, integrase and other enzymes[7].

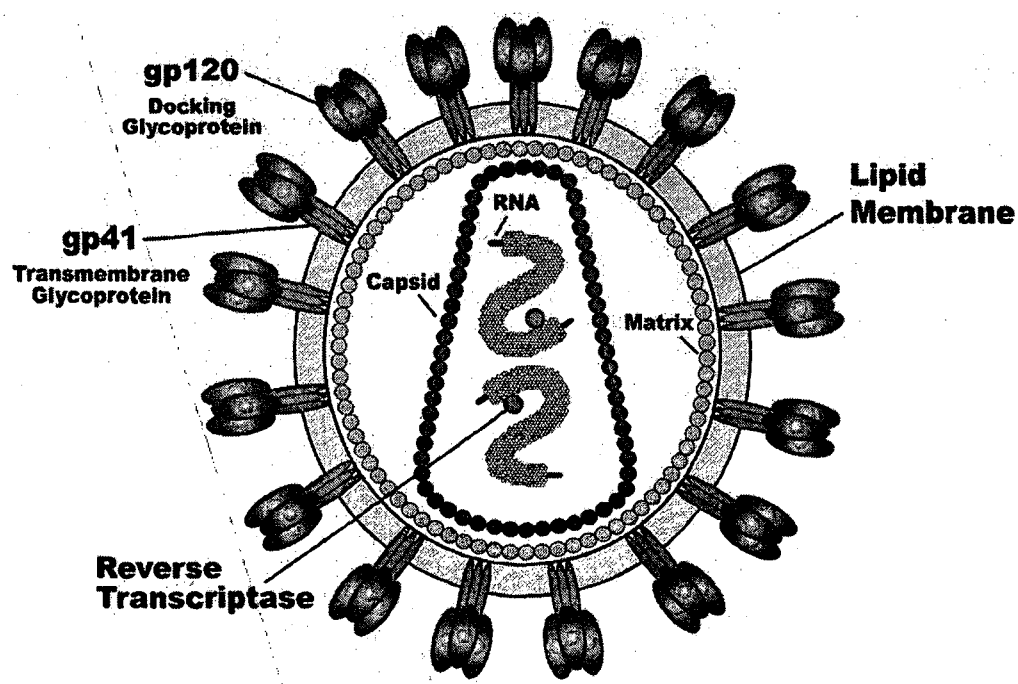


Figure 1.1

Internal Structure of HIV (Source Wikipedia)

1.2.1 Gene map on HIV genome:

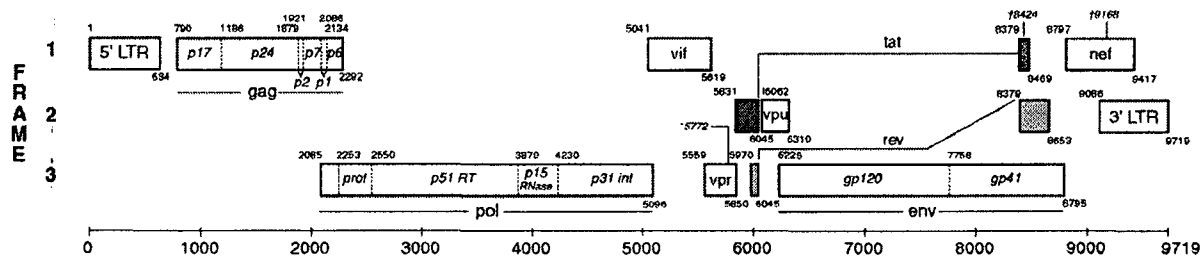


Figure 1.2

Source HIV database tutorial

The key features of strain HXB2 are shown in Figure 1.2, the three frames are shown as rectangles. Generally the ATG is start codon for the gene and it is marked by small number in the upper left corner of each rectangle while the last position of the stop codon is represented by the number in the lower right of the rectangle. The Figure indicates that gag, pol and env protein coding region covers the major part of genome so these three regions capture the unique characteristics of each HIV strain. For the results presented in this thesis,we have used sequence variation of gag-pol coding region for classification purposes. We also carried out classification on the basis of env coding region, which are consistent with the classification based on the gag-pol region.

1.2.2 Genes and gene products:

gag: This protein coding region of the viral genome codes for structural proteins of the virus. Gag is a region that codes for the core structural protein p24 which makes the viral caspid. p6 and p7 are responsible for forming the nucleocaspid. p17 provides the protective matrix for the virus.

pol : The pol coding region codes for the viral enzymes protease, reverse transcriptase and integrase. The reverse transcriptase enzyme is responsible for reverse transcription. It has low level accuracy with an error rate between1/1000 and 1/10000 misincorporation. The protease enzyme is produced as a gag-pol precursor polyprotein. In the presence of integrase, the provirus DNA is integrated into the host genome .

env : This region code for glycoprotein which is the precursor of glycoprotein gp120 and the transmembrane glycoprotein gp41. gp120 provide the binding site for the CD4 receptor. The attachment of gp120 to receptors of CD4 cells is responsible for interaction with gp41. As a consequence of this interaction these three components are fused simultaneously.

1.2.3 Table for main genes with their associated products:

Protein	Gene	Gene products
Structural	gag	Matrix Caspid Nucliocaspid
	env	gp120 gp41
Enzyme	pol	Protease Reverse transcriptase Integrase
Regulatory	tat Rev	Tat Rev
Accessory	vpu vif vpr nef	Vpu Vif Vpr Nef

1.3 Classification structure:

On the basis of sequence variation of different strains of HIV-1 sequences are classified in the following manner.

Groups: The strains of HIV-1 are classified into groups which are M,N, O and P group. Majority of HIV-1 strains belongs to the M group and this group is responsible for the pandemic of HIV.

Subtypes: Within group M there are subtypes A, B, C,D,F, G, H, J and K.

Sub-subtypes: subtype A is classified into A1 and A2 sub-subtypes and F is classified into F1 and F2 sub-subtypes.

Circulating recombination form: The genomes of these type of strains are made up of different segments from more than one distinct subtype. These type of strains are growing due to recombination events, which is one of the normal mechanism of retrovirus replication. They play and important role of increasing viral diversity. For example, the Figure below shows the CRF A/B which is a mixture of subtypes A and B.

CRF03_AB

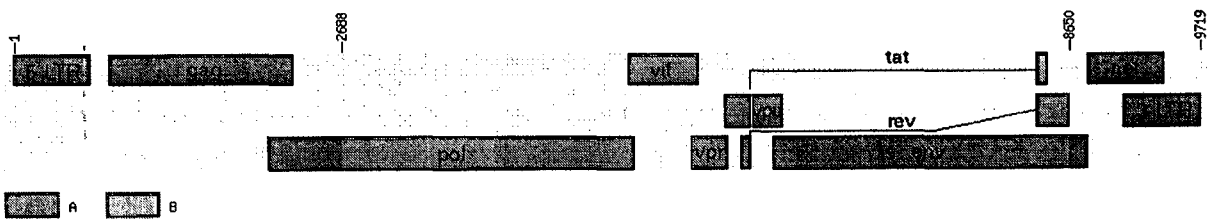


Figure 1.3

Source HIV database tutorial

1.4 Problem for HIV-1 classification: historical aspects:

Initially the classification of different HIV-1 strains was carried out on the basis of sequences obtained from different regions such as Europe North America and Africa . Phylogenetic analysis shows distinct clusters for strains derived from Europe, North America and Africa. Later, when different strains were found from other geographical regions, the classification system based geographical region was not appropriate. However, most classification methods are based on phylogenetic analysis of different strains of HIV1.

First computational approach to classify the different subtypes on the basis of sequence variation of env coding region used phylogenetic analysis [8,10]. Phylogeny provided distinct clusters for A to F subtypes. Eventually gag coding region variation was used for subtyping but E subtype was not classified [9]. E subtype was later found to be the CRF01_AE (a combination of A and E strains) which explains why the gag region was not sufficient to properly classify this subtype[11]. Subtype I was also identified as CRF02_AG [12]. The subtypes A and F are further classified into sub-subtypes A1 , A2 and F1 , F2, and F3 respectively on the basis of differential clustering of env and gag by phylognetic comparison. On the basis of complete genome analysis sub-subtype F3 classified in to new group K [13] . All the existing strains of HIV type 1 subtype grouped into three clades which were M group, N group and O group [5].

Most of the HIV strains were formed the same clusters when different genomic region were analyzed. However it was found earlier that branching of phylogenetic tree of some strains was unresolved for different parts of its genome. This fact suggests that these strains are product of recombination events. These type are known as circulating recombination form(CRF) of strains[5]. Till now, number of computational methods exist for detection of different subtype and also to analyze recombinants of different types of CRF strains. We emphasize that classification of HIV -1 stains is complex issue. A fast and efficient method which can be provide the recognition of existing sequences as well the sequences appearing in the future is required.. The method must be easily adapted to work with not only whole genome sequences but also with parts of a genome. It should be effective in classification even if a small number of known sequences of a particular subtype is available. In our knowledge HIV-1 subtyping program(STAR)[14], which provide the classification on the basis of sequence variation of the segment of protease (PR) and reverse transcriptase (RT), is similar to our

work. STAR is based on position specific scoring matrices for each subtype which derived from multiple sequence alignment. It classify different subtype of pure strains of HIV-1 with high degree of accuracy. This accuracy lies between 90% to 98%. Our method classify pure strains of HIV-1 with 100% accuracy and also extensible for classification on the basis of sequence variation of different coding regions. Therefore accuracy and flexibility of our method to manage the different segments of genome is advantageous over STAR.

1.5 HIV-1 classification using profile hidden Markov models:

Profile hidden Markov models have proved to be extremely successful in detecting homologous protein and nucleic acid sequences from large data sets[15]. In our study, we demonstrate that pHMMs provide an efficient, reliable and robust method for classifying HIV strains into different subtypes with a high degree of accuracy. First we will show that the standard classification method of pHMMs using bit score can be successfully employed for classification of all pure subtypes (except B and D) with perfect accuracy. We find that the standard method cannot accurately discriminate between some sequences belonging to subtypes B and D. These shortcoming are completely removed by using an improved method using positive and negative pHMMs which is successfully applied for detection of different subtypes with 100% accuracy. The improved method utilizes the Z score, which is characterized as a measure of position specific features presented in query sequence which are unique to a given subtype, to recognition that subtype. We have presented the classification performance of different subtypes on the basis of the the Z scores of each query sequence which belong to the test data set. On the basis of performance of distribution of Z scores of query sequences of test set over different subtypes we are able to find the fixed threshold (Z=0) for distinguishing between true positives and true-negatives. This threshold (Z=0) is the same for all subtypes. Hence our improved method has the capability to accurately assigning the subtype to strains which are sequenced in future.

We also extended our method for detection of recombinants for each query sequence of CRF strains. This method was able to identify the subtype composition of a CRF strain with a high degree of accuracy. However, we found that the performance of detection of different subtype can be affected by the length of the segment of the subtype in the genomic region (example: gag-pol) used for building the profiles.

Chapter 2

Methods and Material

2.1 Data acquisition :

All the available 1511 pure genome sequences (excluding CRFs) belonging to the M group available in the Los Alamos database were downloaded in fasta format . The required annotated sequences of the gag-pol and env segments of the coding region for each of the subtypes were downloaded in fasta format for building the profile HMMs using appropriately chosen positive and negative training sets. All the CRF genome sequences (i.e. sequences containing a mixture of two or more subtypes) were also downloaded to determine their subtype content.

Used database : For requirement of data, the Los Alamos HIV-1 sequence database stores updated sequences and which are easily fetched under required criteria . This database also stores sequences of annotated coding regions.

2.2 Hidden Markov model :

A Hidden Markov Model (HMM) is completely specified by the following quantities (Rabiner,1989) [16]:

1. The set of states $\{S_1, S_2, ..., S_n\}$ and state q_t at time t.

2. The set of output alphabets $\{v_1, v_2, \ldots v_m\}$ and the output O_t at time t.
3. The probability π_i of being in state S_i at time t = 0 .
4. The transition probability matrix $A=[a_{ij}]$ where
   $a_{ij} = P[q_j, t+1 \mid q_i, t]$.
5. The emission probability matrix $B=[b_j(k)]$, where $b_j(k)$ is the probability that system emits the output v_k given that it is in jth state

2.3 Profile hidden Markov model :

Hidden Markov Models have been successfully applied in speech recognition . Anders Krogh, David Haussler, and co-workers adopted HMMs to find patterns in protein sequences and also introduced profile HMMs (pHMMs)[15]. Profile Hidden Markov Models are statistical models that are derived from multiple sequence alignment of homologous sequences of nucleic acids or proteins. It captures the position specific features that are present in the multiple sequence alignment for the corresponding protein or nucleic acid family. Such features are characteristics of the homologous sequences.

A score is generated with help of its parameters for a query sequence . To discriminate between the homologous and non-homologous sequences, a threshold is decided on the basis of which optimum classification is possible.

A pHMM is a particular example of HMM that consists three types of states corresponding to each column position in a multiple sequence alignment . Suppose a multiple sequence alignment has L columns. Then the match state , insert state and delete state corresponding to pth position (column) in multiple sequence alignment are denoted by M_p , I_P and D_p respectively where $1<=p<=n$. These states emits symbols with a certain emission probability distribution over the alphabets on which the pHMM is defined. In case of nucleic acid sequences the set of alphabet is {A,T,G,C} and in case of protein sequences ,the set of alphabets is the set of 20 biologically encoded amino acids. B and I_0 is defined as the match state and insert state at initial position and E is match state at the final position. The begin states and end states are called dummy state since these states do not emit any symbol. In the case of pHMMs there are three transition are possible from any state (except E state) to another . Self transitions are possible for only insert states which indicates that the multiple insertion is only allow.

The states and the possible transition among different states both determine the topology of the model. All the transition probabilities among the states and emission probabilities over the alphabets are known as parameters of profile HMM. The Baum-Welch algorithm is used to estimate the parameters of the model .

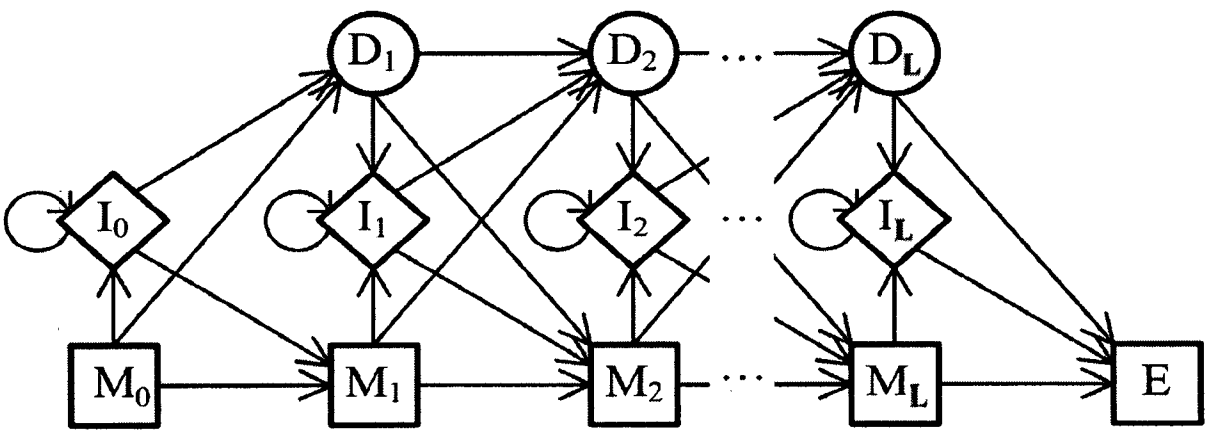


Figure 2.1

2.4 Profile HMM scoring :

pHMMs have the capability to recognize the membership of a particular query sequence in a family on the basis of significant matches of the query sequence to the profile HMM . There are many type of scores which are used for membership determination[17].

2.4.1 Log likelihood score :

The log likelihood (LL) score is defined as logarithm of probability of the sequence given the model (M).

$$LL(\text{sequence}) = \log P(\text{sequence}|M)$$

The LL score depends on the length of the sequence. From the classification point of view, it is not

appropriate to use this scoring method since it does not yield a simple threshold for discriminating between true positives and true negatives. There is a way to remove this shortcoming by dividing the LL score by the sequence length. However, this does not solve the problem completely since for all cases the dependency between the LL score and the sequence length is not linear.

2..4.2 Log-odd score relative to random model :

The log-odds score relative to a random model (R) is a popular measure for determining whether a query belongs to the family described by the pHMM. The log-odd score (S) is defined as the log of the probability of obtaining the sequence from a profile HMM divided by the probability of pobtaining the same sequence from a random model .

$$S = \log_2 \frac{P(\text{sequence}|M)}{P(\text{sequence}|R)}$$

In HMMER, the random model (R) is a simple one-state profile HMM that emits the alphabets of sequences with independently and uniform probability distribution over its alphabet. So this term P(sequence|R) is dependent on only length of sequence[18].

HMMER provides the log-odds score for the complete sequence which is also known as the bit score. The bit score is independent of the size of the sequence database, and depends only on the profile HMM and the query sequence. For classification purposes this bit score can used and a cutoff for the score is set in such a way that optimum classification is possible.

2.5 Selection of coding region for training the pHMMs:

The segment of gag-pol coding region is cover nearly half portion of the whole genome of HIV-1. So, we have taken nucleotide sequences variation of gag-pol coding region for classification of different subtypes. However different subtypes are detected on the basis of sequence variation of env coding region.

2.6 Multiple sequence alignment (MSA) :

MUSCLE provide improved alignment accuracy compared with other currently available MSA programs[19]. MUSCLE is freely available at http://www.drive5.com/muscle. We used MUSCLE 3.7 (Edgar, 2004) to generate the MSA that was used to build the profile HMM for the sequences making up the training set.

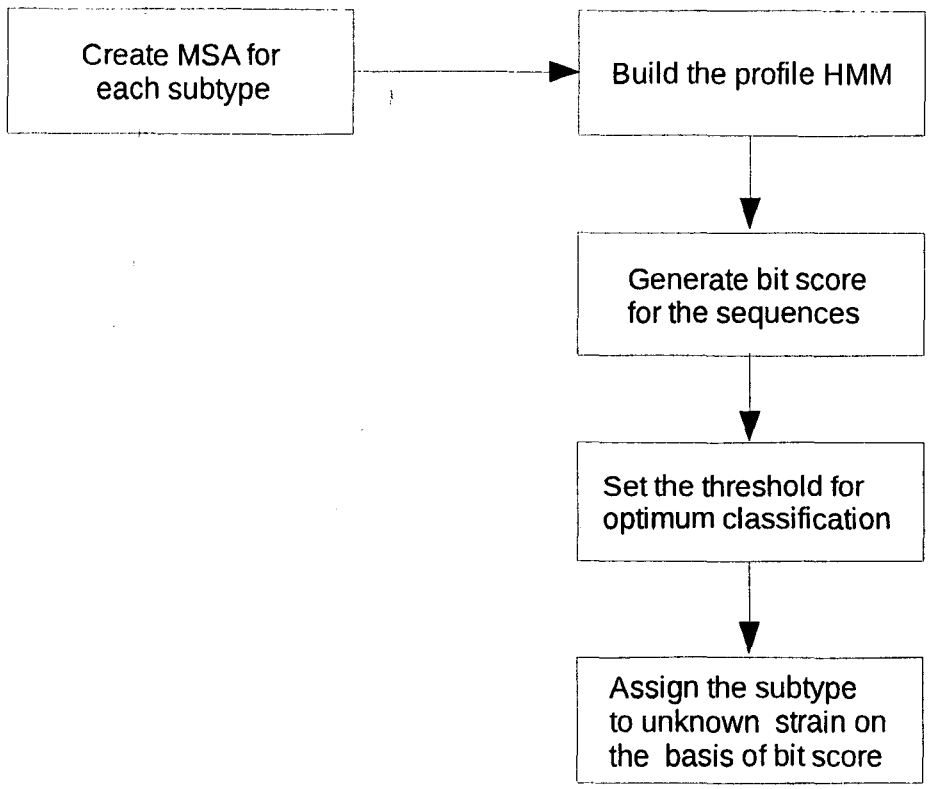
2.7 Selection of positive and negative training set:

Positive training set: We have randomly taken six distinct sequences of gag-pol each subtype forbuilding the positive pHMM to corresponding subtype. The accession no of for each subtype is listed in table 2.11.1 .

Negative training set: Negative training set for a given subtype contains two distinct sequences of gag-pol coding region except that subtype. The accession number which are used for negative training set are listed in table 2.11.2 .

2.8 Standard method for classifying HIV-1 sequences to appropriate subtypes

We created multiple sequence alignment of each subtype from sequences containing the gag-pol segment. A pHMM was generated from each of the multiple sequence alignments by using program hmmbuild program from the HMMER package. The score (S) for each 1511 genome sequences was generated by the hmmsearch program in HMMER .We developed a classification method on the basis of log odd or bit score (S) . In this method, all the sequences have positive scores . We set a threshold to discriminate between true positives and true negatives. This threshold is dependent on each subtype classification. For all the subtypes except B and D we are able to find an unambiguous threshold which allowed for accurate discrimination of true positives and true negatives. The diagram below shows the flowchart-of the classification method. thing this method, we were able to classify with a high level of sensitivity and specificity sequences of all subtypes other than subtypes B and D.



Flow diagram of classification by the standard method of profile HMM

Classification result when profile HMM is build by gag-pol coding segment of B subtype and visualization the bit score for each subtype green line represent lowest bit score of B subtype strains and blue line stands for greatest bit score of strains except B .But in this case there are some some strains which are fall within this threshold . Now if we ignore threshold as blue line then there are some false positive but at this case error is low.

Classification result when profile HMM is build by gag-pol coding segment of D subtype and visualization the bit score for each subtype green line represent lowest bit score of D subtype strains and blue line stands for greatest bit score of strains except D . In this manner shows very poor classification only two members are correctly classified .

In this way we analyze that profile HMM can not apply for correct assignment of all the subtype when only single profile HMM which is created by only positive training set . We remove this shortcoming by selection of two profile HMM corresponding to classification of each subtype.

2.9 Improved method for subtype classification:

For explanation of correct classification method we have to define following terms :

Set of positive training sequences : For purpose of classification of each subtype of the M group there is associated set which contains some sequences of the gag-pol coding region of corresponding subtype, the accession numbers of which are give in Table 1. These sequences are randomly chosen without redundancy and number of sequences that make up the positive training set for each subtype is six.

Set of negative training sequences: The negative training set contains two sequences of gag-pol coding region for all subtypes excepting the one used to create the positive training set. The accession number of all the sequences used to create the negative training set are given in Table 2.

Positive profile HMM: For each subtype there is a set of positive training sequences we use hmmbuild program in HMMER and create positive profile HMM for the associated subtype .

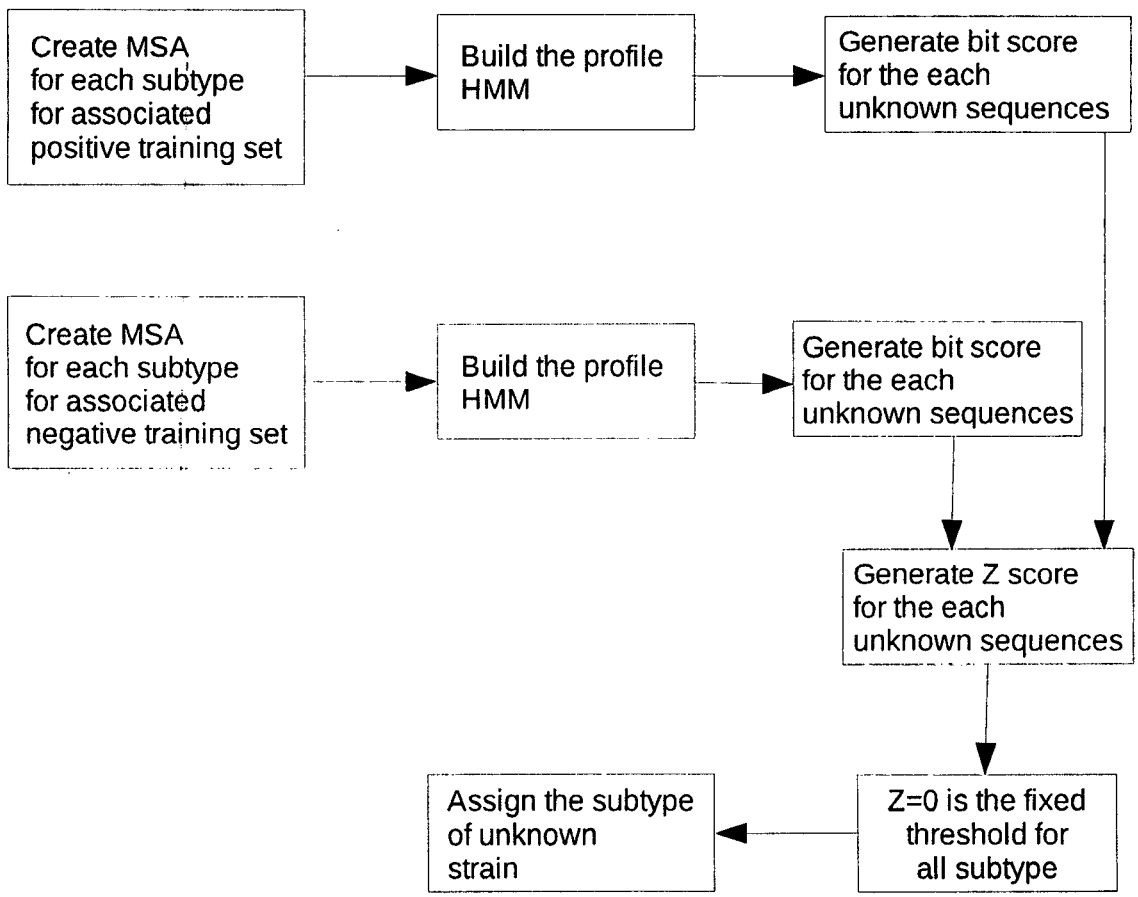
Negative profile HMM: For each subtype there is a set of a associated negative training sequences we use hmmbuild program in HMMER and create negative profile HMM for the associated subtype .

Classifying subtype: Classification method for a subtype and subtype of sequences which are contain in corresponding set of positive training sequences are same. This subtype is known as classifying subtype.

Z score: Z score of a query sequence can be defined as the bit score of the query sequence which is generated by the corresponding positive profile HMM minus the bit score of the query sequence which is generated by the corresponding negative profile HMM . This Z score is measure of subtype specific signal and is an effective measure for accurately classifying the query sequence.

Multiple Sequence alignment is created using MUSCLE package for required sequence file to build profile HMM. Positive and negative Profile HMMs provide the direction to create the method to successfully detect the given subtype from unknown subtype of pure strains belonging to test set. The program hmmsearch of HMMER is used to generate the bit score of all 1511 genome sequences for given subtype of positive and negative profile HMMs. Then we compute the Z score using a perl script for each query sequence. This method assign the subtype of query sequence if its Z score is positive which is generated by the positive and negative pHMMs of that subtype. In this case strains of A, B, C, F and G subtypes are detected with 100% accuracy. But problem is arisen in detection of D subtype strains. The discrimination power is increased if we take more sequences of those subtypes in negative training set, which are more similar to subtype of strains which belong to the corresponding positive training set, instead of taking equal number of sequences of different subtypes of strains. So in order to remove this problem, we have taken six sequences of gag-pol coding region of B subtype strains instead of two sequences. Strains of D subtype are accurately detected by considering the sequence variation of env coding region.

In order to sub-typing of A there are the there we choose two sequence from A subtype two from A1 sub-subtype and two from A2 sub-subtype. For F subtype required six sequences are chosen from F1 sub type. All sequences of F1 subtype classified properly but F2 sub-subtype does not belong to positive value But three member closer to zero threshold .



Flow diagram of improved method

2.10 Method for analysis of CRFs and unclassified sequences :

pHMMs can be successfully applied to identify the different subtypes which make up a CRF strains. Profile HMMs have been built by positive and negative training set, which provide the method to successfully detect given subtype as a recombinant in unknown CRF strains. The discrimination power is increased if we take more sequences of those subtypes in negative training set, which are more similar to subtype of strains which belong to the corresponding positive training set, instead of taking equal number of sequences of different subtypes of strains. As a result the number sequences in negative training set are increased. For clear discrimination, we increased the number of sequences of positive training set instead of six. The number of sequences in positive and negative training set are given in tables 2.12.1 and 2.12.2 . Suppose we want to determine whether the subtype D is present in a CRF strain. We define two thresholds : (i) Tp which is near the minimum Z score of all the D subtype pure sequences and (ii) Tn which is near maximum Z score of all the pure sequences that *do not* belong to the D subtype (i.e. sequences which make up the negative training set when the positive training set is made out of sequences belonging to the D subtype.) If CRF strain returns a Z score greater than Tn and less than Tp then it is predicted that some portion of the gag-pol segment of D subtype is present in the CRF strain. If the CRF strains returns a Z score greater than Tp then the gag-pol segment of this strain can be predicted to consist of either pure D subtype or a mixture in which a large fraction of the gag-pol region is of D subtype . The method can be repeated using positive training sets made up of different subtypes to determine the subtype composition of the gag-pol region of the CRF. The subtype composition of other segments of the CRF can also be determined in a similar manner. This method can fail to detect a particular subtype in a CRF if the segment of that subtype is too small to contain signatures specific to that subtype, that have been captured by the pHMM.
We generalize the method for detection of query sequence which have recombinant of a given subtype by the following assumptions .

1. If the Z score for a given query sequence lies between Tp and Tn then the gag-pol segment of the query sequence contains some portion of the subtype from which the positive training set was constructed.

2. If the Z score for a given query sequence is lower than Tn then gag-pol coding region of this query sequence is not made up with a gag-pol segment of the corresponding subtype.
3. If the Z score for a given query sequence is greater than Tp then this gag-pol coding region of query sequence is made up with pure sequence of gag-pol segment of the corresponding subtype.

If the CRF strains is made of two or more different subtype of recombinant strains with significant length. These type of strains return Z score for each corresponding subtype detection under the above assumptions. We will see how this method detects the CRF strains which are made up with B and F subtypes.

2.11 Accession numbers of training set for detection of pure sequences:

2.11.1 Positive training set:

S.N.	A subtype	B subtype	C subtype	D subtype	F subtype	G subtype
1	AM000053	AB097870	AF110963	AY773340	GQ290462	AB287003
2	AM000054	AB286956	AF286228	DQ054367	DQ979025	FJ389364
3	AB098330	AB287370	AF110974	AB485650	AB485659	AB485663
4	AB253421	AB289589	AB254150	AJ519489	AJ249238	AY586548
5	AF286237	AB428560	AB097871	A14116	AB480300	AF423760
6	AF286238	AB480698	AB485645	U88822	DQ189088	AY612637

2.11.2 Negative training set :

S.N.	Subtype	Accession
1	A1	AB253422
2	A2	AF286238
3	B	A04321
4	B	AB287372
5	C	AB254141
6	C	AB485645
7	D	A34828
8	D	AY773340
9	F1	AB485656
10	F2	AJ249237
11	G	AB485662
12	G	AY586548
13	H	AF005496
14	H	FJ711703
15	J	AF082394
16	J	GU237072
17	K	AJ249235
18	K	AJ249239

2.11.3 Negative training set used for D subtype classification :

S.N.	Subtype	Accession
1	A1	AB253422
2	A2	AF286238
3	B	AB565496
4	B	AF042102
5	B	AB480696
6	B	AF538305
7	B	AB485642
8	B	AF049495
9	C	AB254141
10	C	AB485645
11	F1	AB485656
12	F2	AJ249237
13	G	AB485662
14	G	AY586548
15	H	AF005496
16	H	FJ711703
17	J	AF082394
18	J	GU237072
19	K	AJ249235
20	K	AJ249239

2.12 Number sequences used in training sets for detection of different types of CRF strains :

2.12.1 Number of sequences of positive training sets used in detection of different types of CRF strains:

Subtypes	Number of sequences
A	6
B	31
C	6
D	14
F	10
G	6

2.12.2 Number of sequences of negative training sets used in detection of different types of CRF strains:

A subtype		B subtype		C subtype		D subtype		F subtype		G subtype	
subtype	number	subtype	number	subtype	number	subtype	number	subtype	number	subtype	number
A	0	A	2	A	2	A	2	A	2	A	6
B	2	B	0	B	2	B	31	B	2	B	2
C	2	C	2	C	0	C	2	C	2	C	2
D	2	D	16	D	2	D	0	D	2	D	2
F	2	F	2	F	2	F	2	F	0	F	2
G	6	G	2	G	2	G	2	G	2	G	0
H	2	H	2	H	2	H	2	H	2	H	2
J	2	J	2	J	2	J	2	J	2	J	2
K	2	K	2	K	2	K	2	K	2	K	2

Chapter 3

Results For Pure Sequences

In this chapter, we present the results of classification of query sequences belonging to an unknown subtype (excluding CRF's) of the group M using pHMM's for each subtype that have been constructed from training sequences. Our results indicate that pHMM's provide a powerful method for accurately classifying query sequences into appropriate subtypes with a high degree of sensitivity and specificity.

3.1 Performance of pHMM as a classifier:

To determine the effectiveness of our pHMM models in accurately classifying query sequences, it is necessary to determine the sensititivity and specifity of the model. Suppose a model has the capability to predict either membership or non-membership in a given subtype, for unknown data belonging to the test set. The result of that prediction can fall into any one of the four possible categories.

1. The given query sequence is correctly predicted to belong to the associated subtype: True positive(TP) .

2. The given query sequence is incorrectly predicted as not belonging to a particular subtype: False Negative(FN) .

3. The given query sequence is correctly predicted as not belonging to a particular subtype: True Negative(TN) .

4. The given query sequence is incorrectly predicted as belonging to a particular subtype: False Positive(FP).

The sensitivity and specificity may be used to measure the performance of classifier and is defined as

$$\text{Sensitivity} = TP/(TP+FN)$$

$$\text{Specificity} = TP/(TP+FP)$$

3.1.1 Receiver operating curve :

Receiver-Operator Characteristic (ROC) curves (Sensitivity i.e true positive rate v/s 1 - specificity i.e. false positive rate) for classification of each subtype indicate the discriminating potential of the profile HMM.

The table below gives the list of abbreviation that are used in plots showing the bit-scores of sequences belonging to the different subtypes.

Abbreviation of level	Subtype
A1	Strains of A1 subsubtype
A	Strains of A subtype
A2	Strains of A2 subsubtype
B	Strains of B subtype
C	Strains of C subtype
D	Strains of D subtypes
F1	Strains of F1 subsubtype
F2	Strains of F2 subsubtype
G	Strains of G subtype
U	Unclassified strains

3.2 Performance of the standard method of profile HMM :

A pHMM is constructed for each HIV-1 subtype using a positive training set only. It is then used to determine the bit score of all sequences which may or may not belong to that particular subtype. which is used to classify only the sequences of corresponding subtype from the test data set. The construction of the profile HMM each subtype is described in the chapter "Method". For a given threshold score,which is arbitrary chosen between Tp and Tn, for a given subtype, the model can be used to determine whether a query sequence can be considered to be a member of the corresponding subtype on the basis of its bit score. A query sequence return a bit score, which is generated by pHMM trained with multiple alignment of sequences of given subtype, greater than the threshold then we assign it to corresponding subtype. Using this method, sequences belonging to A, C, F and G subtypes are accurately classified by the associated profile HMM. The Figure below shows the bit score of all sequences when the training set is constructed using sequences belonging to the C subtype. The plot shows a clear demarcation between the bit-scores of those sequences that are members of the C subtype and those that aren't. Classification with 100% accuracy is obtained by choosing a threshold that lies between the bit score values Tp and Tn indicated by the green and blue lines respectively.

3.2.1 Classification of C subtype by standard profile HMM :

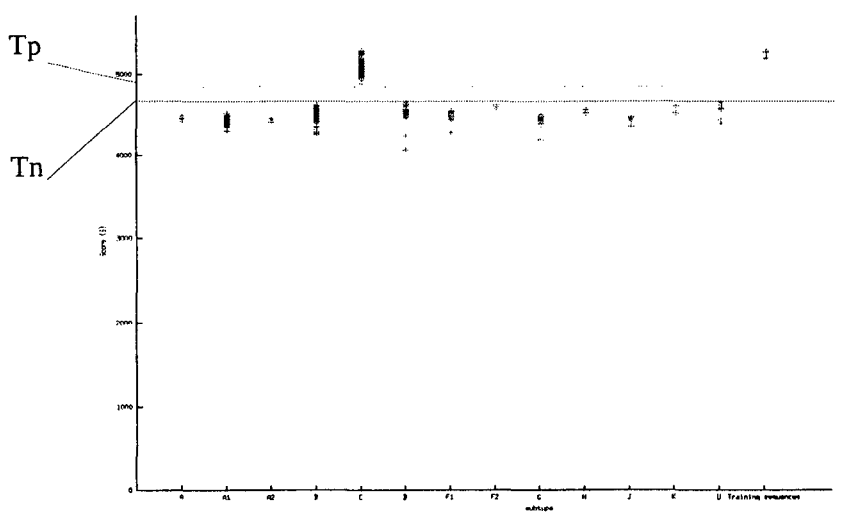


Figure 3.1

Similar results are obtained for classification of sequences that belong to other subtypes of the M group. However, the standard method was unable to accurately discriminate between some sequences belonging to the B and D subtypes. The reason for this problem is that the genetic distances, between B subtype and D subtype strains, are less in comparison to other subtype.

Classification of B subtype by standard profile HMM :

3.2.2 Classification of B subtype by standard profile HMM:

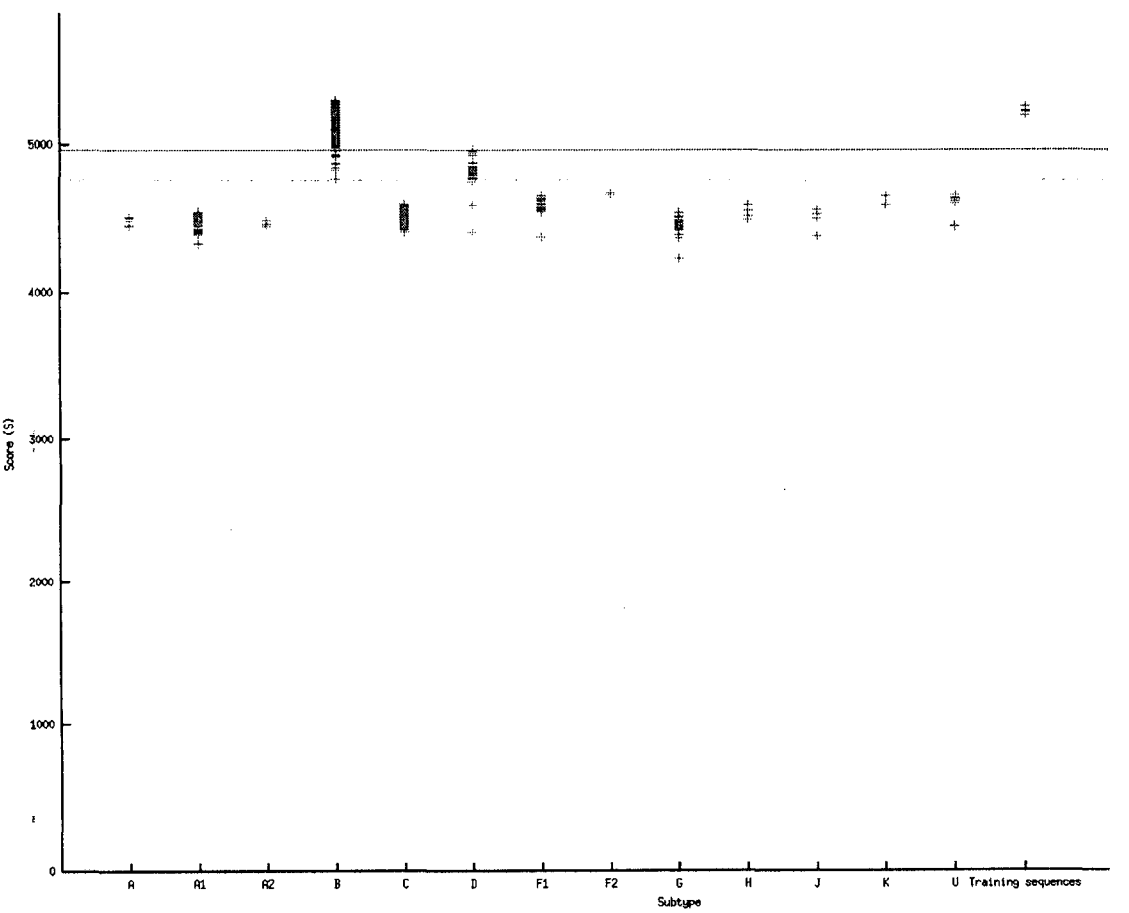


Figure 3.2

Classification result when profile HMM is build by MSA of six sequences of gag-pol coding segment of B subtype.

This plot shows the bit scores for each subtype; the green line represents the lowest bit score of the B subtype strains and the blue line stands for the greatest bit score of sequences from all strains excepting B. In this case there are some some strains which fall between these two lines. Therefore we are not able to find any threshold which gives accurate classification of B subtype.

A similar problem is observed when we attempt to accurately classify sequences belonging to the D subtype by constructing a training set with six sequences of the gag-pol coding region belonging to the D subtype. The green line and blue line stands for lowest bit score of D subtype strains and the highest bit score of sequences belonging to all strains except D. In this case D subtype strains are poorly classified, since only two members are correctly identified when the blue line is chosen as the threshold.

3.2.3 Classification of D subtype by standard profile HMM:

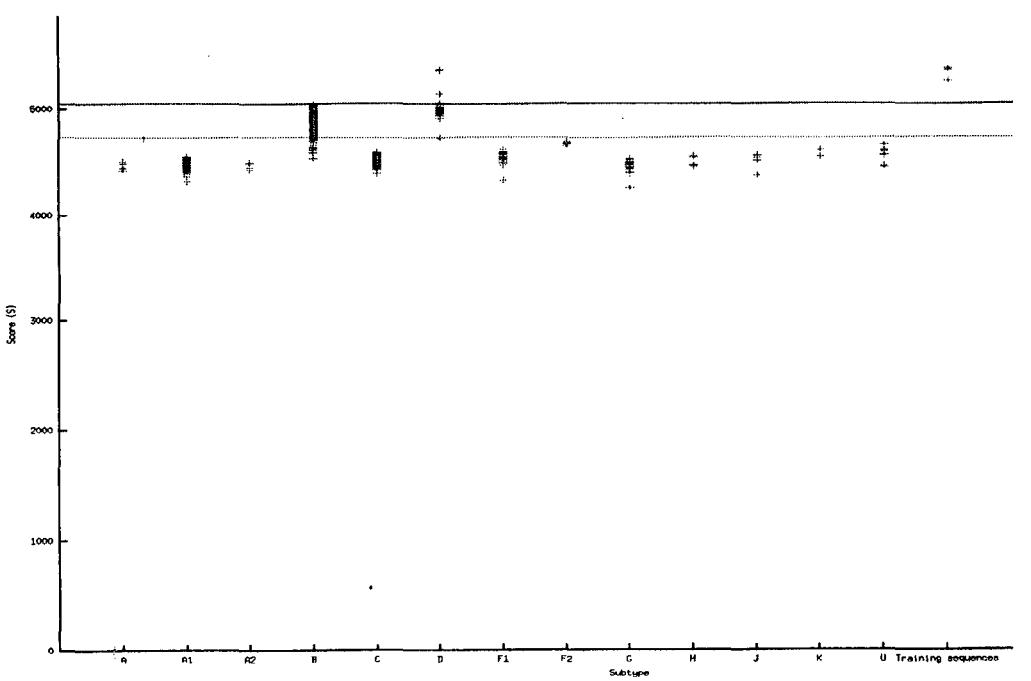


Figure 3.3

There are two shortcoming have arisen in classification by standard method :

1. The strains of a subtype, which(subtype) is lesser dissimilar to another subtype than pairwise dissimilarity of possible combination of different subtypes, are not accurately detected.
2. The thresholds are not independent from different subtype. Fixed threshold remove arbitrariness between Tp and Tn. We have the need for certain assumption which decide the fixed threshold that could robustly determine the subtype of strains which are sequenced in future.

The drawbacks of the standard method discussed above, suggests the need for developing a method in which the threshold can be unambiguously fixed for all subtype classification and which is better able to discriminate between sequences that are closely related but belong to different subtypes, as in the case of sequences belonging to subtypes B and D.

3.3 Results from the improved method for subtype classification :

We introduced an improved method using pHMMs which could accurately and robustly identify the subtypes of unknown strains with fixed threshold. For a given subtype there are associated positive pHMM and a negative pHMM, as discussed in detail in the chapter "Methods". The Z score of a query sequences for a given subtype measures position specific features in the query sequence which are unique to the corresponding subtype. So performance of the method for different subtypes on the basis of the the Z scores of each query sequence belonging to the test data set, have increased. This method assign the subtype of query sequence if its Z score is positive otherwise this query sequence does not belongs to that subtype. We will showed the different plot for classification performance of our method. In each plot the threshold (Z=0) is fixed i.e independent from different subtype detection. This method has improved ability to discriminate the strains belonging to more similar subtypes. Strains belonging to B subtype are detected with 100% accuracy. However problem is arisen in detection of D subtype strains but this is removed by taking six sequences of B subtype in stead of two sequences in negative training set.

We present a graph in which the X-axis lists the name of each subtype and the Y-axis gives the Z-scores of strains belonging to each subtype. In case of accurate classification of given subtype, all the strains of which belong to that subtype have Z scores greater than zero. subtype U corresponds to unclassified strains . These strains do not get positive Z scores for any classification that uses a positive training set constructed from one of the known subtypes. However in case of classification of A subtype one member unclassified strain (AY046058) closer to threshold which shows it have more similarity with A subtype.

The Z score, which is generated by the method for classification of strains of a given subtype by taking equal number of sequences of each different subtype in negative training set, interpret the information of pairwise distances between the cluster of strains of that subtype and the clusters of strains belonging to different subtypes. In this case we have take two sequences of each subtype in negative training set. In Contrary, this information can not be interpreted when number of sequences of a those subtype have been increased which are more similar to the subtype of corresponding positive training set. But it provides the direction for clear discrimination between strains of closely related subtype. In order to discriminate D subtype of strains we can take thirty one sequences of B subtype instead of two sequences of each subtypes except B in its negative training set . In this case we have taken forteen sequences of D subtype in positive training set.

Histogram plot of 1511 strains vs Z score for a classification of a each subtype are drawn to show the distribution number of sequences with different Z score. Visualization of histogram plots also represent the the information of distribution of different strains of positive subtype in its cluster. In case of higher number of strains belonging to positive subtype more strains of the positive subtype are close to the center of its cluster.

Receiver operating characteristic (ROC) curves for each subtype shown below indicates that the performance of classification by this method is 100% accurate.

3.3.1 Classification result for A subtype :

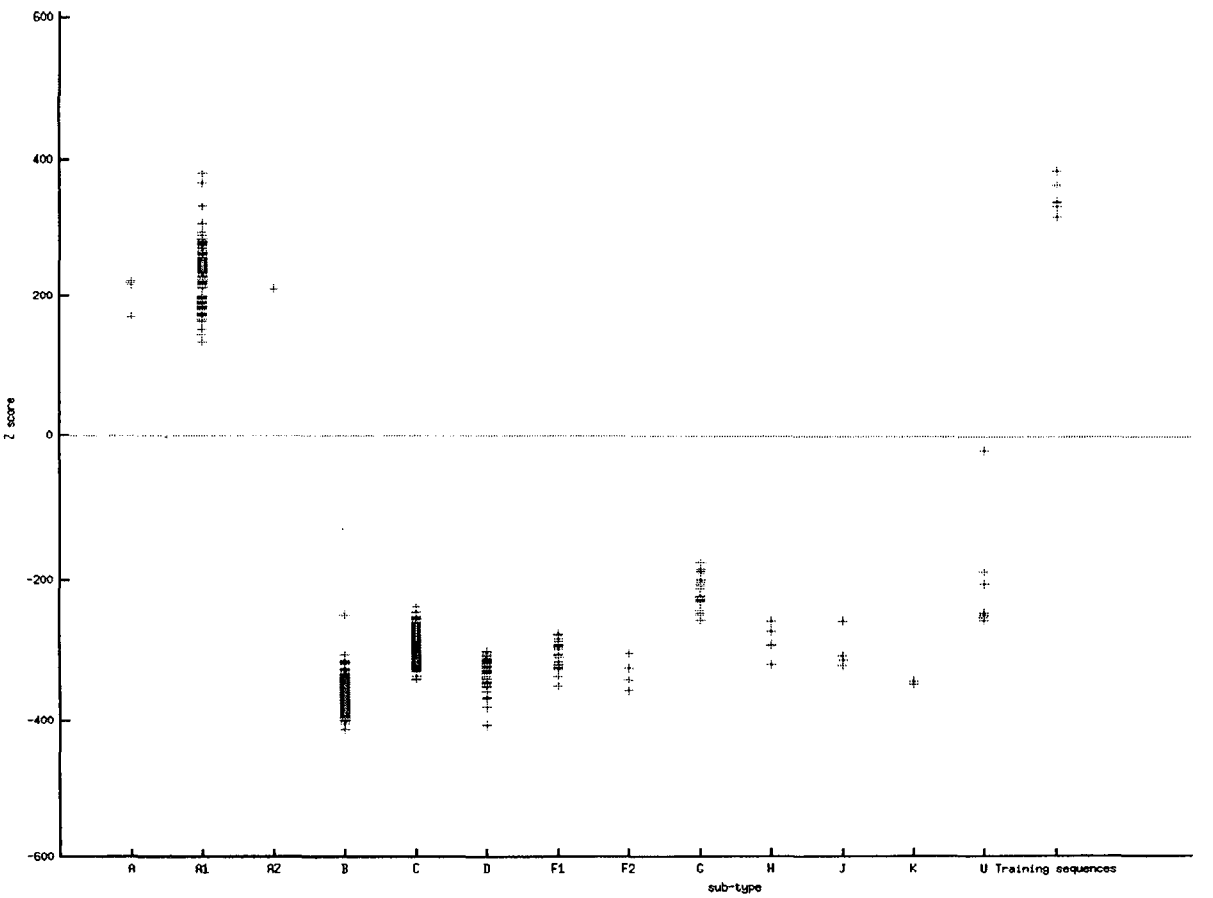


Figure 3.4

Distribution of Z scores of different subtypes of HIV -1 strains of M group. Level U stands for unclassified strains. Cluster of A subtype strains are more similar to the cluster of G subtype in comparison to other subtypes. The clusters A subtype strains are less dissimilar to B and D subtype strains in comparison to other subtypes. One unclassified strain(AY046058) is closer to A subtype in terms of its Z-score. However this strains has correctly not been assigned to the A subtype.

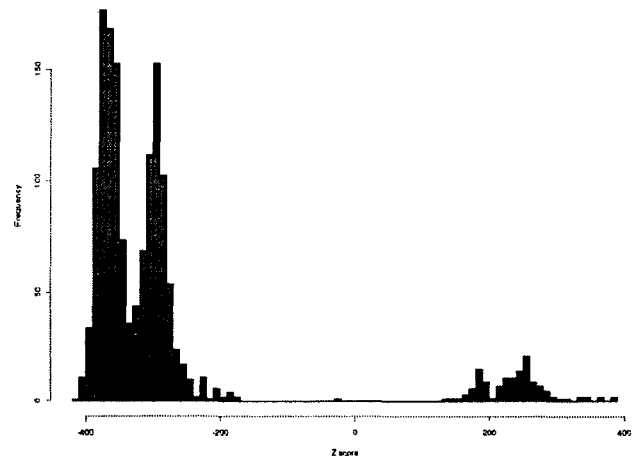


Figure 3.5

Frequency distribution of 1511 genome sequences of HIV-1 strains of Z score which is generated to accurately identify members belonging to the A subtype.

3.3.1.1 Receiver operating curve for A subtype classification :

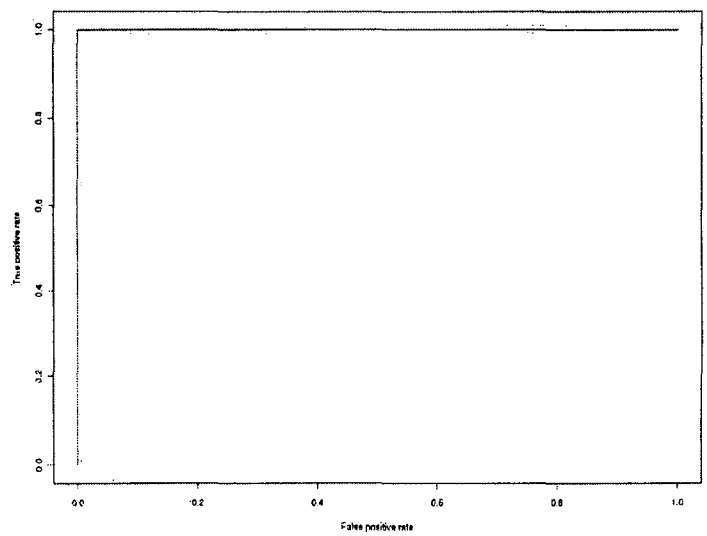


Figure 3.6

ROC analysis of improved method for A subtype classification. This plot shows profile HMMs provide the 100% classification accuracy for B subtype.

3.3.2 Classification result for B subtype:

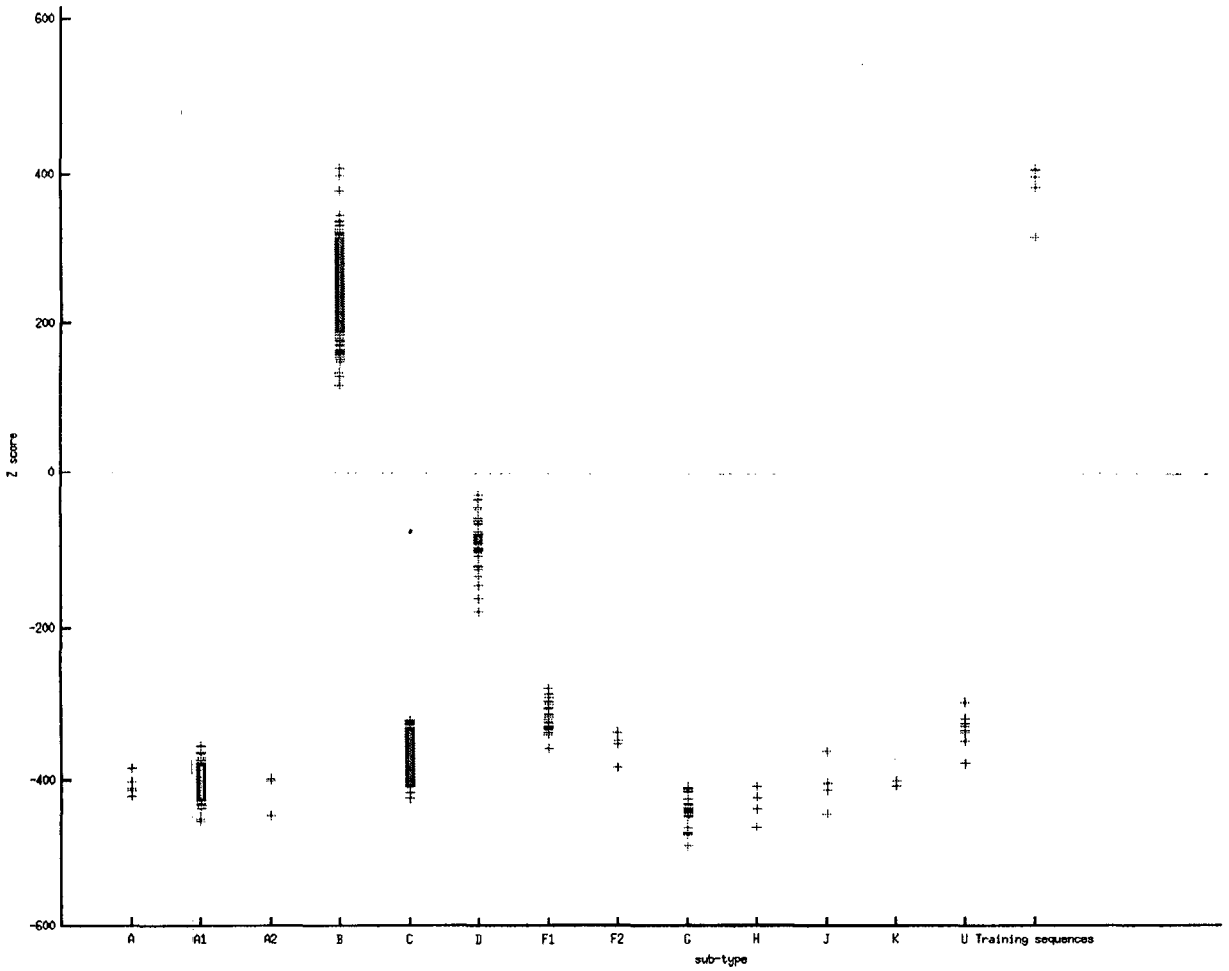


Figure 3.7

Distribution of Z score of different subtypes of HIV -1 for B subtype of classification when positive pHMM is trained by multiple alignment of six sequences of gag-pol coding region. This plot indicate the cluster of B subtype strains are more similar to cluster of D subtype in comparison to other subtypes . The cluster of B subtype strains is less dissimilar to A and G subtype strains in comparison to other subtypes. Cluster of B subtype strains are also less dissimilar to F subtype strains however this dissimilarity is lesser than of cluster of strains of D subtype .

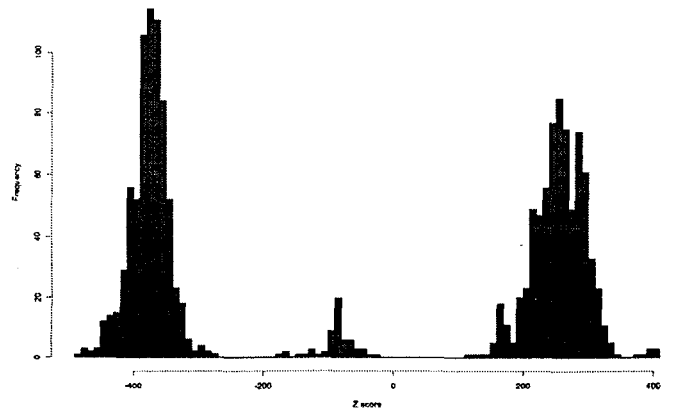


Figure 3.8

Histogram plot of 1511 genome sequences of HIV-1 strains vs Z score which is generated to classify B subtype .This plot shows the number of sequences with different Z scores. There are three clusters right one is for B subtype middle one for D subtype and left one is for all the subtypes except B and D . The cluster of strains of B subtype shows the number of stains are higher in neighborhood of center of its cluster.

3.3.2.1 Receiver operating curve for B subtype classification :

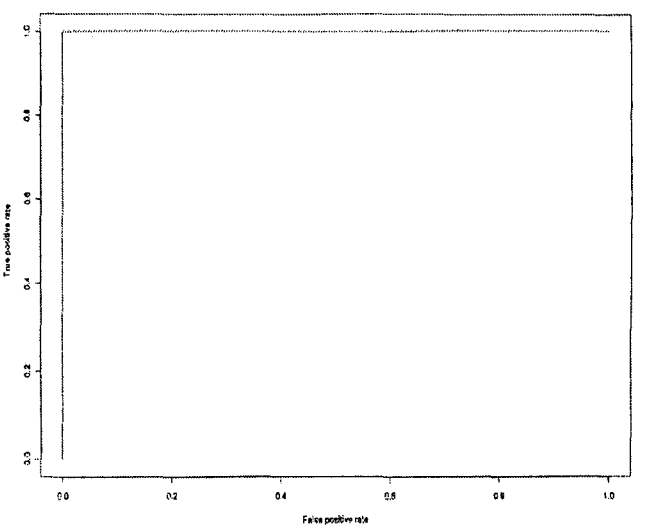


Figure 3.9

ROC analysis for strains of B subtype of classification using improved method. This plot shows profile HMMs provide the 100% classification accuracy for B subtype.

3.3.3 Classification result for C subtype:

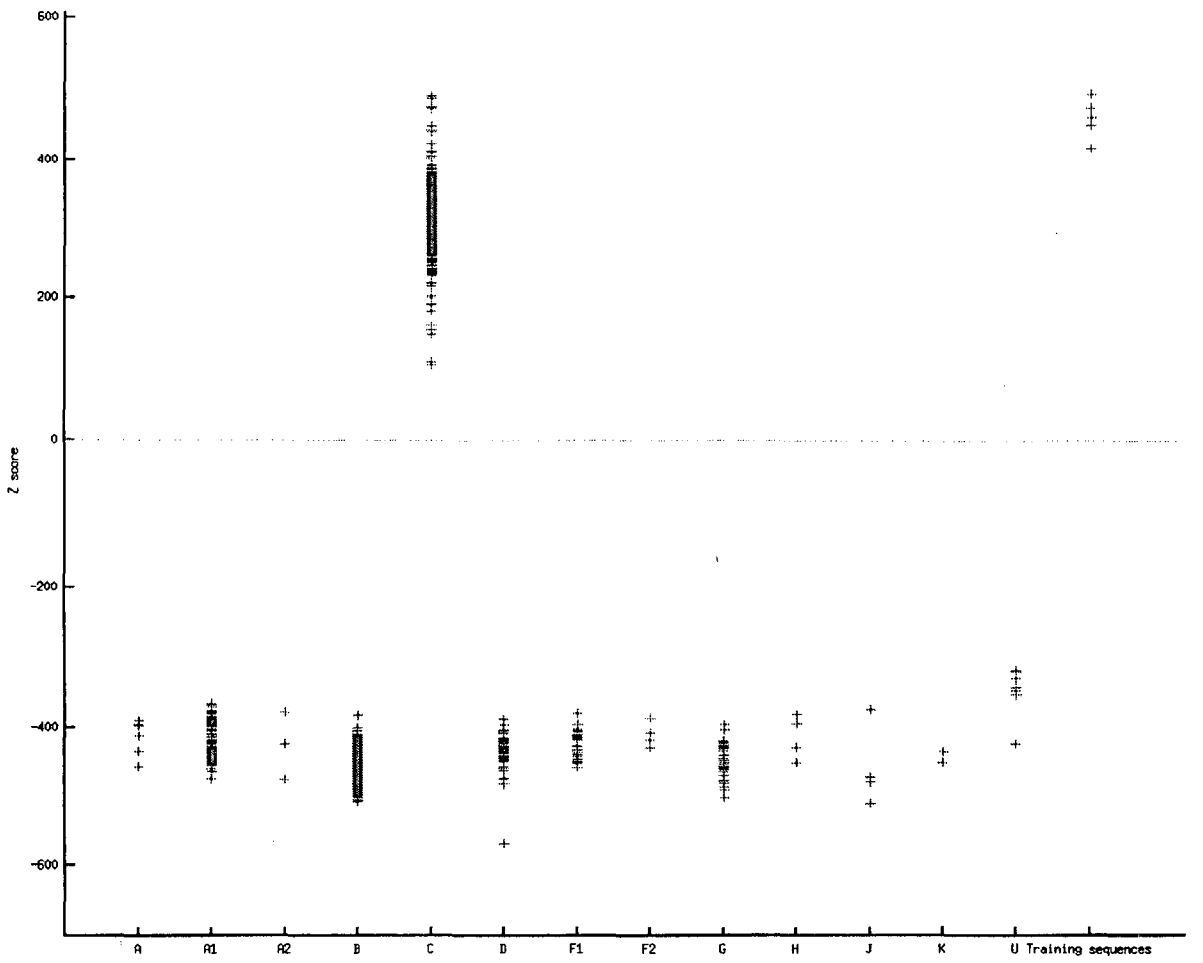


Figure 3.10

Distribution of Z score of different subtypes of HIV -1 . In this case positive profile HMM is built by multiple alignment of six sequences of the gag-pol coding region of C subtype strains . The plot shows the C subtype strains are detected with 100% accuracy.

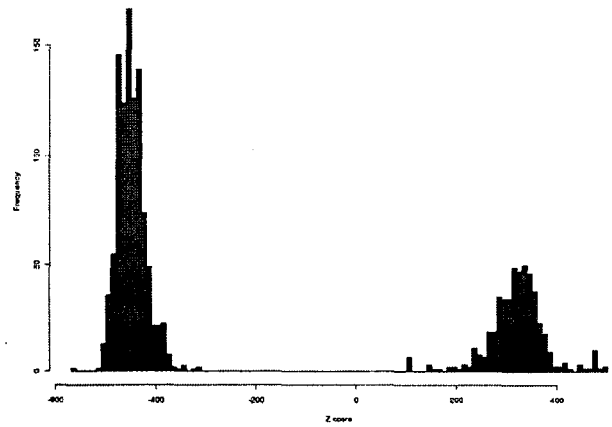


Figure 3.11

Histogram plot of 1511 genome sequences vs Z score which is generated for classification of C subtype. This plot shows the number of sequences with different Z score which is generated for C subtype classification. The cluster of strains of B subtype shows the number of stains are higher in neighborhood of center of its cluster.

3.3.3.1 Receiver operating curve for C subtype classification :

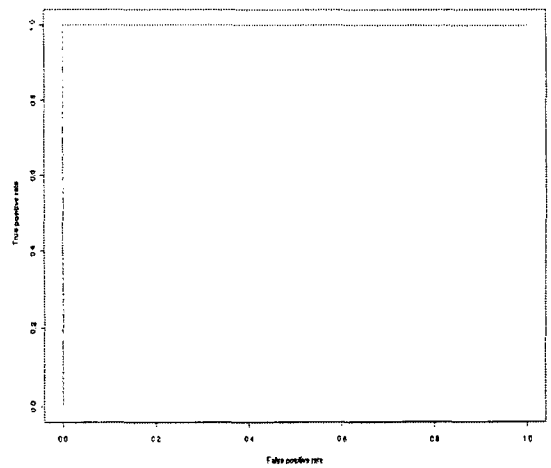


Figure 3.12

ROC analysis of classification of strains of C subtype using improved method. ROC curve is plot between true positive rate vs false positive rate. Area below the curve is maximum so there are not false positives or false negatives. This plot shows profile HMMs provide the 100% classification accuracy C subtype.

3.3.4 Sub-typing result for D subtype:

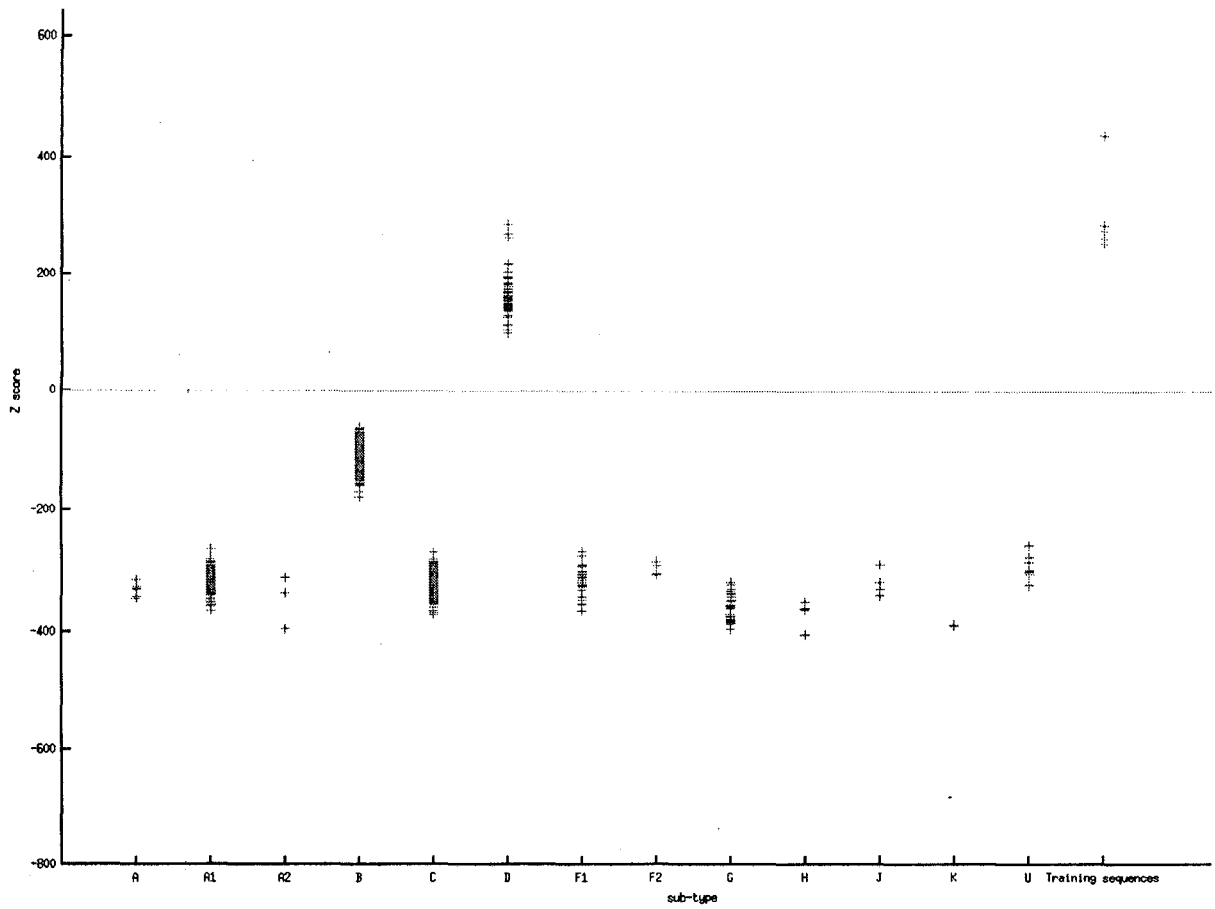


Figure 3.13

Distribution of Z score of different subtypes of HIV -1 for classification implementation of D subtype. The strains of B subtype are very close to threshold which indicates cluster of D subtype strains are more similar to cluster of B subtype in comparison to all other subtypes .

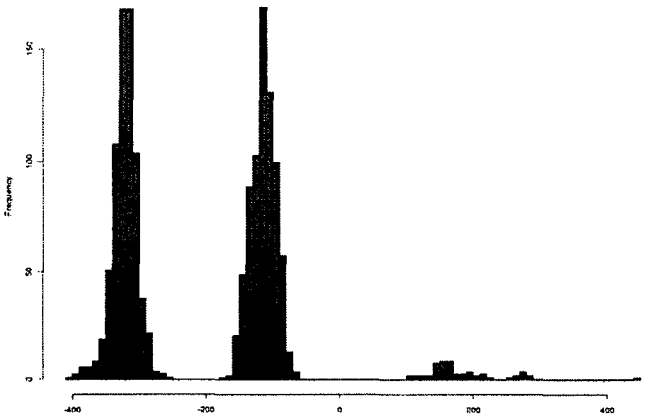


Figure 3.14

Histogram plot of 1511 genome sequences vs Z score which is generated for classification of D subtype. This plot shows the number of sequences with different Z score which is generated for D subtype classification.

3.3.4.1 Receiver operating curve for D subtype classification:

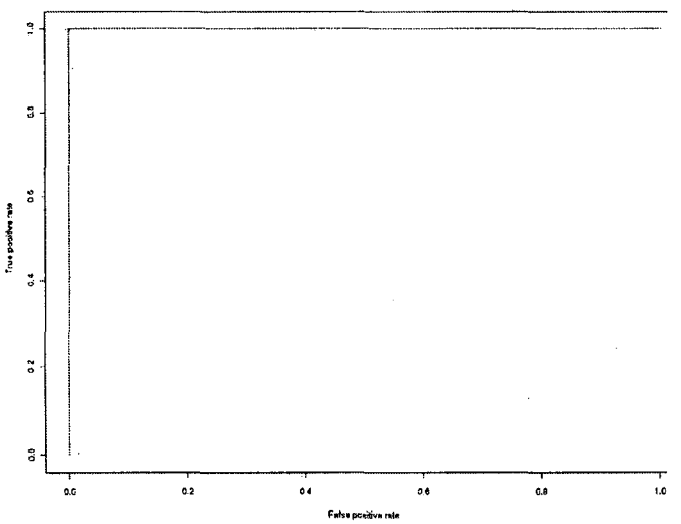


Figure 3.15

ROC analysis of D subtype classification using improved method. ROC curve with true positive rate vs false positive rate. Area under the curve is maximum so there is no False positive and also there is not false negative. This plot shows profile HMMs provide the 100% classification accuracy for D subtype.

3.3.5 Classification result for F subtype:

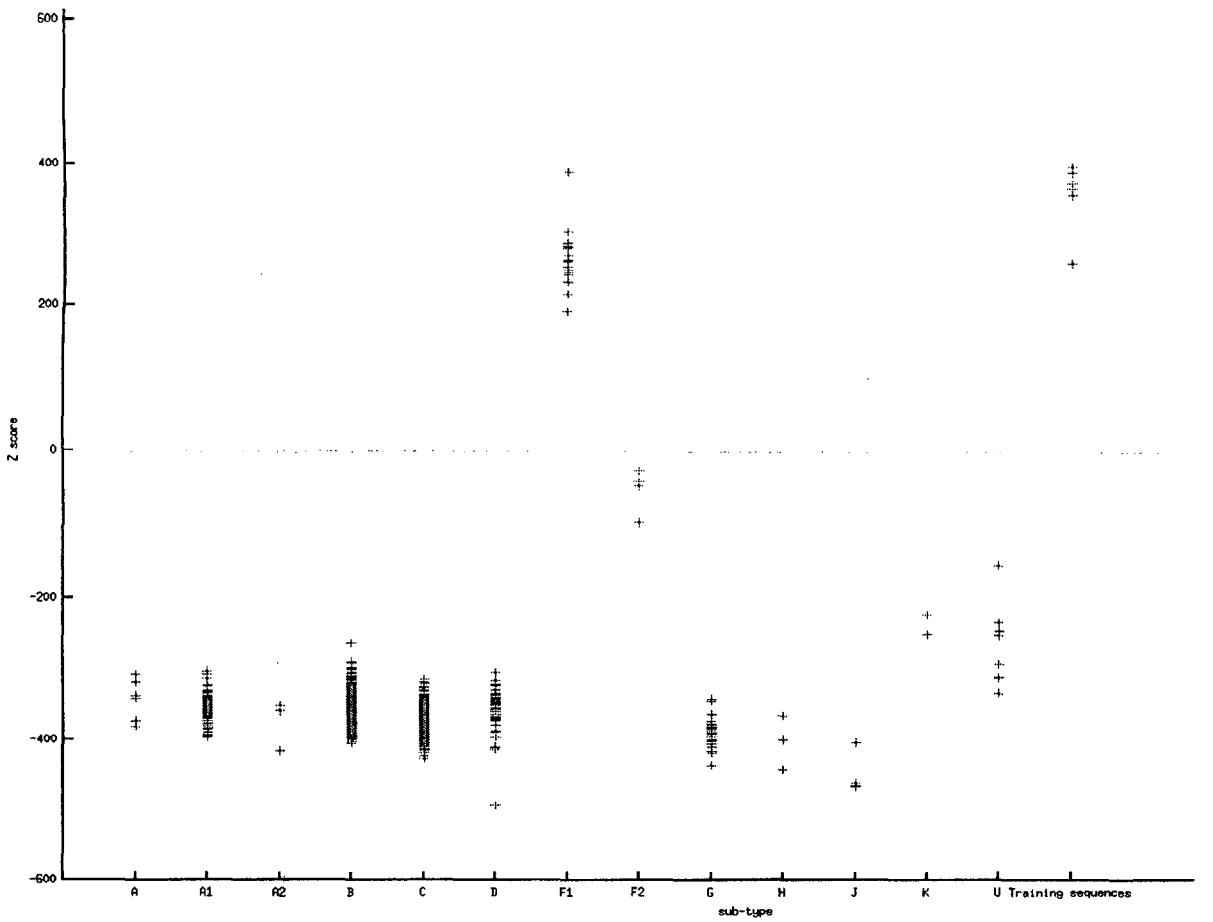


Figure 3.16

Distribution of Z scores of different subtypes of HIV-1 for classification of F. In this case positive profile HMM is built by multiple alignment of six sequences of F1 subsubtype. F2 subsubtype strains are closer to the threshold. One unclassified strains (FJ388921) is close to threshold which indicate this strain has more signal that are similar to the F1 subsubtype .

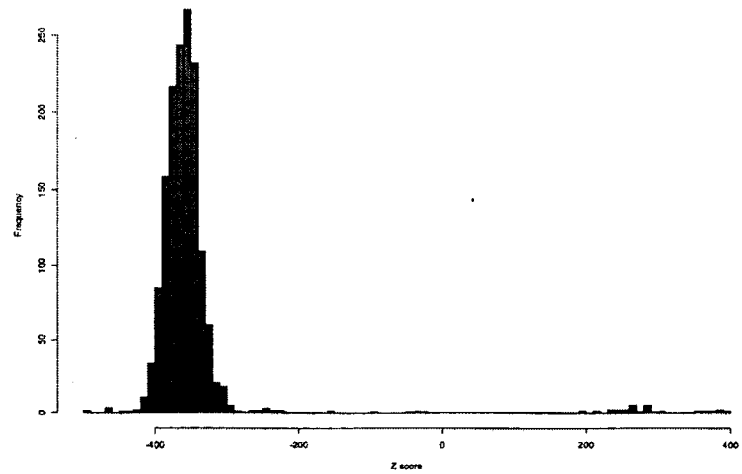


Figure 3.17

Histogram plot of 1511 genome sequences, on x axis we take Z score which is generated for classification of F1 subsubtype . This plot shows the number of sequences with different Z score which is generated for F1 subsubtype classification.

3.3.6.1 Receiver operating curve for F1 sub-subtype classification :

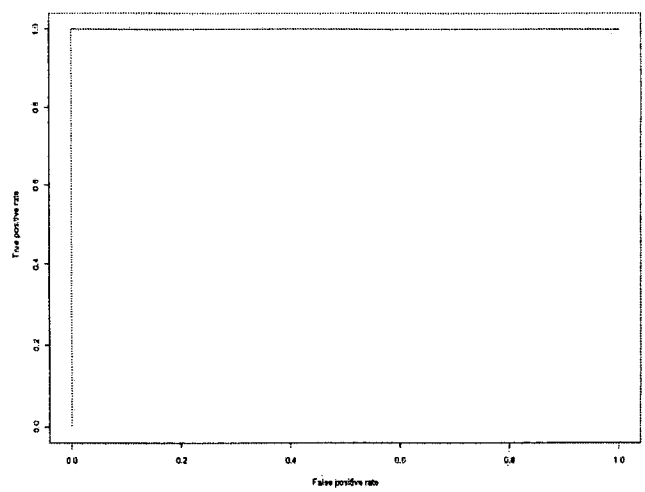


Figure 3.18

Receiver operating curve analysis of F1 subsubtype classification using improved method. This plot shows profile HMMs provide the 100% classification accuracy for F1 subsubtype.

3.3.7 Classification result for G subtype:

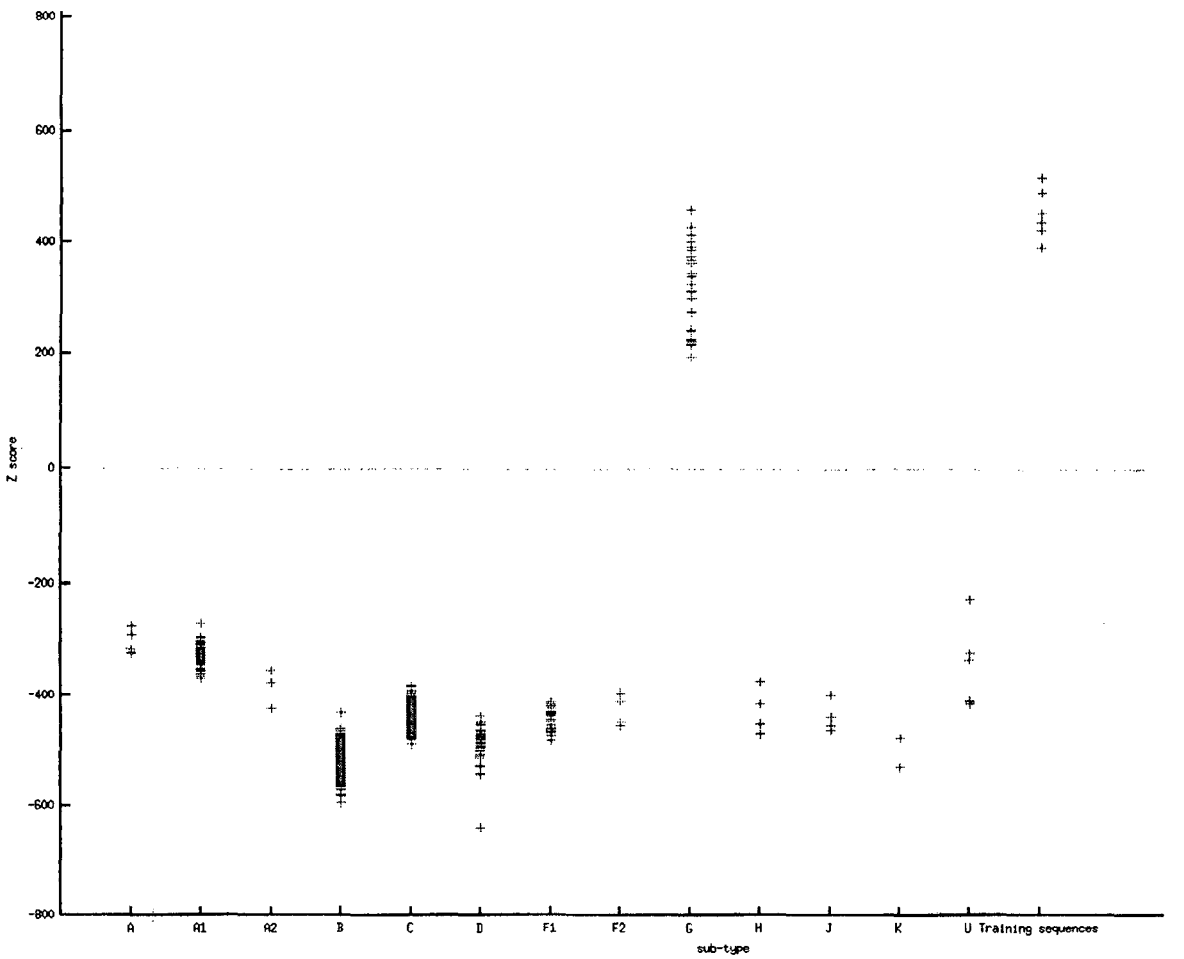


Figure 3.19

Distribution of Z score of different subtypes of HIV -1. In this case positive profile HMM is built by multiple alignment of six sequences of gag-pol coding region of G subtype strains. This fig shows that the genetic distance between cluster of different strains of G subtype strains and cluster of different strains of A subtype strains is less dissimilar in comparison to its distances from cluster of strains of any other subtypes .

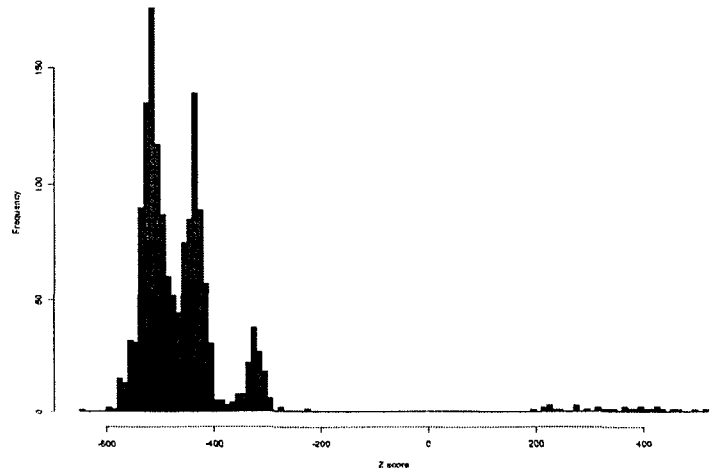


Figure 3.21

Histogram shows the number of sequences vs Z score which is generated for G subtype .This plot shows the number of sequences with different Z score which is generated for G subtype classification.

3.3.7.1 Receiver operating curve for G subtype classification:

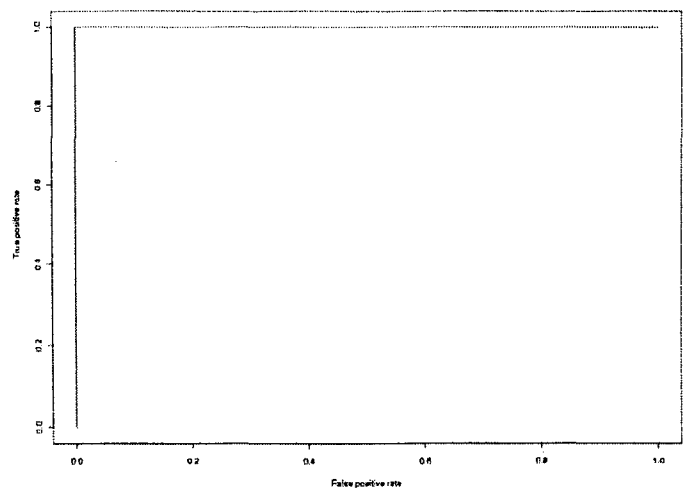


Figure 3.22

ROC curve analysis of G subtype classification using improved method. This curve is drawn between true positive rate vs false positive rate. Area under the curve is maximum so there is not False positive and false negative. This plot shows profile HMMs provide the 100% classification accuracy for G subtype.

Chapter4

Sequence Analysis of Strains of Circulating Recombination Form

In this chapter we will present the performance of the improved method, when it is used to detect the different subtypes which make up the CRF strains. We have described so far the classification of different subtypes of pure sequences and how performance of the method is improved, when we implemented our method on the basis of Z score, in the chapter "Classification of Pure Sequences". We will see in this chapter how the Z score coupled with two thresholds Tp and Tn, allow for identification of the subtypes which make up a CRF strain. The results on test data set, which contains 2033 genome sequences of pure as well as different CRF strains, are shown for validation of the method.

4.1 The performance of the method :

Profile HMMs have been built by positive and negative training set, which provide the direction to create the method to successfully detect given subtype as a recombinant in unknown CRF strains. The discrimination power is increased if we take more sequences of those subtypes in negative training set, which are more similar to subtype of strains which belong to the corresponding positive training set, instead of taking equal number of sequences of different subtypes of strains. As a result the number sequences in negative training set are increased. For clear discrimination, we increased the number of

sequences of positive training set instead of six. For example sequences of B subtype and sequences of D subtype are more similar, in order to detect sequences of different strains, which are made up as some part of D subtype, we take 31 sequences of B and two sequences of each subtype except B and D in negative training set of D subtype. The number of sequences in negative training set are forty five instead of fourteen to balance thresholds we have taken fourteen sequences of D subtype in positive training set instead of six. As a consequence it distribution of maximum Z score of each different pure sequence except B are approximately same and it help to determine threshold Tn.

We present the plots for detection of each subtype, which has been involved to recombine with significant length, in genetic recombination of different CRF strains of test data set. X axis of this plot have been labeled with notation of different subtypes and different notations of CRF strains. Y axis represent Z score for detection of given subtype. These Z scores have been generated using the method discussed in chapter 3. There are 48 levels for different types of CRF strains. Two thresholds are shown one is Tp and other is Tn as defined in chapter 3. The following assumptions have been taken for the recognition of query sequence and the presence as some fraction of given subtype in the query sequence.

1. If the Z score for a given query sequence lies between Tp and Tn then the gag-pol segment of the query sequence contains some portion of the subtype from which the positive training set was constructed. .
2. If the Z score for a given query sequence is lower than Tn then gag-pol coding region of this query sequence is not made up with a gag-pol segment of the corresponding subtype.
3. If the Z score for a given query sequence is greater than Tp then this gag-pol coding region of query sequence is made up with pure sequence of gag-pol segment of the corresponding subtype.

If a given CRF strains is made of two or more different subtype of strains with significant length then this type of strains return Z score for each corresponding subtypes. We will see how this method detects the CRF strains which are made up with B and F subtypes. However, problems arise only in those cases when either recombination event occurs out side of the gag pol segment or in case a particular subtype is present only as a very short sequence in the gag-pol region.

4.2 Results for detection of CRF strains which contains B subtype strains :

In this section we show how each query CRF sequence that has sections of B subtype in the gag-pol region can be identified using our pHMM method. There are five plots to present the distribution of Z scores, which are generated for detection of B subtype.

Distribution of Z scores with different pure subtypes and unclassified category. The two thresholds Tp and Tn are marked in this graph. The genetic distance between cluster of strains of B subtype and the cluster of strains of D subtype are less in comparison to the distance between other subtypes (chapter 3 Figure). To improve discrimination between B and D, we have taken twenty two sequences of gag-pol coding region of D subtype and two sequences of each subtype other than D in negative training set of the B subtype. Thirty one sequences of gagpol coding region of B subtype are taken in positive training set. As a result this plot shows all the maximum Z scores of each different subtype are nearly same which define the threshold Tn for B subtype.

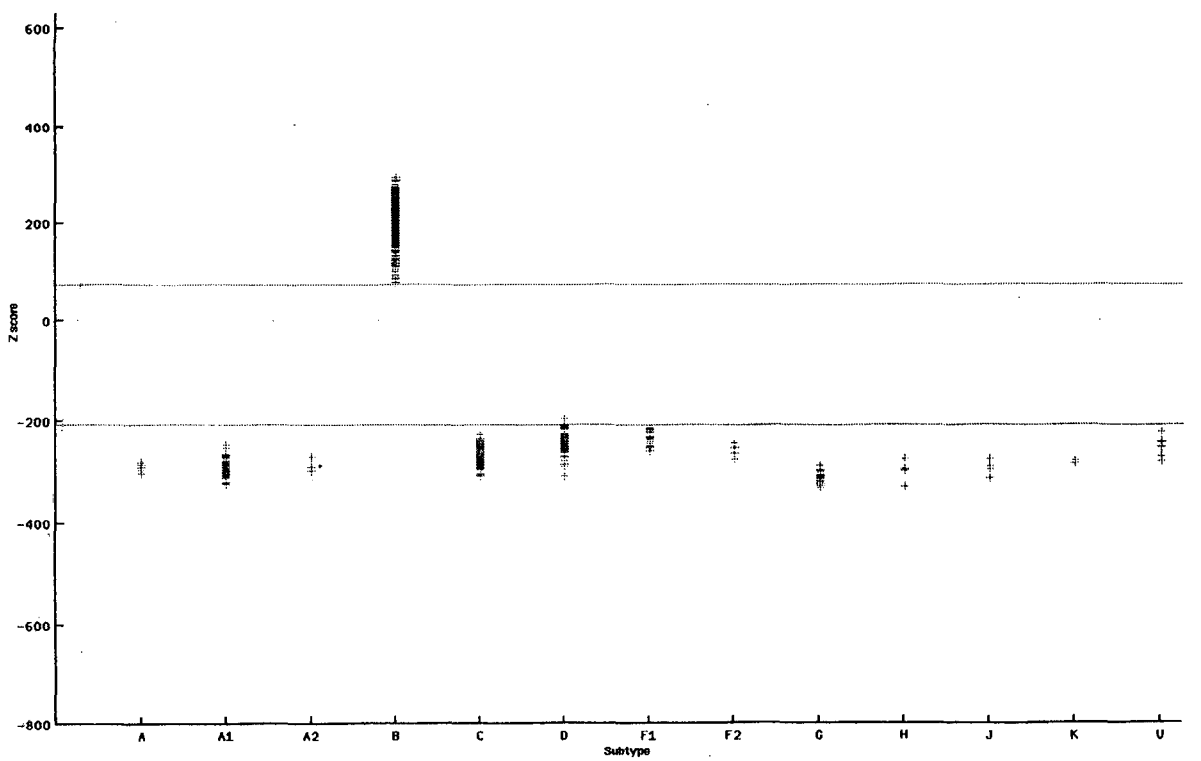


Figure (4.1)

In this plot, level of different types of CRF strains from CRF01 to CRF13 are shown to detect those strains which are recombination as a fraction of gag-pol segment of B subtype. All the strains of type CRF03_AB return Z score between Tp and Tn, therefore this prediction under the our assumption shows that B subtype involve as a part to recombine in its composition. Each strain of types CRF07_BC and CRF12_BF are made up from B subtype. The strains of type CRF08_BC are very close to the threshold Tn which indicate that these types of strains are made up with minimal length of gag-pol coding region of B subtype segment.

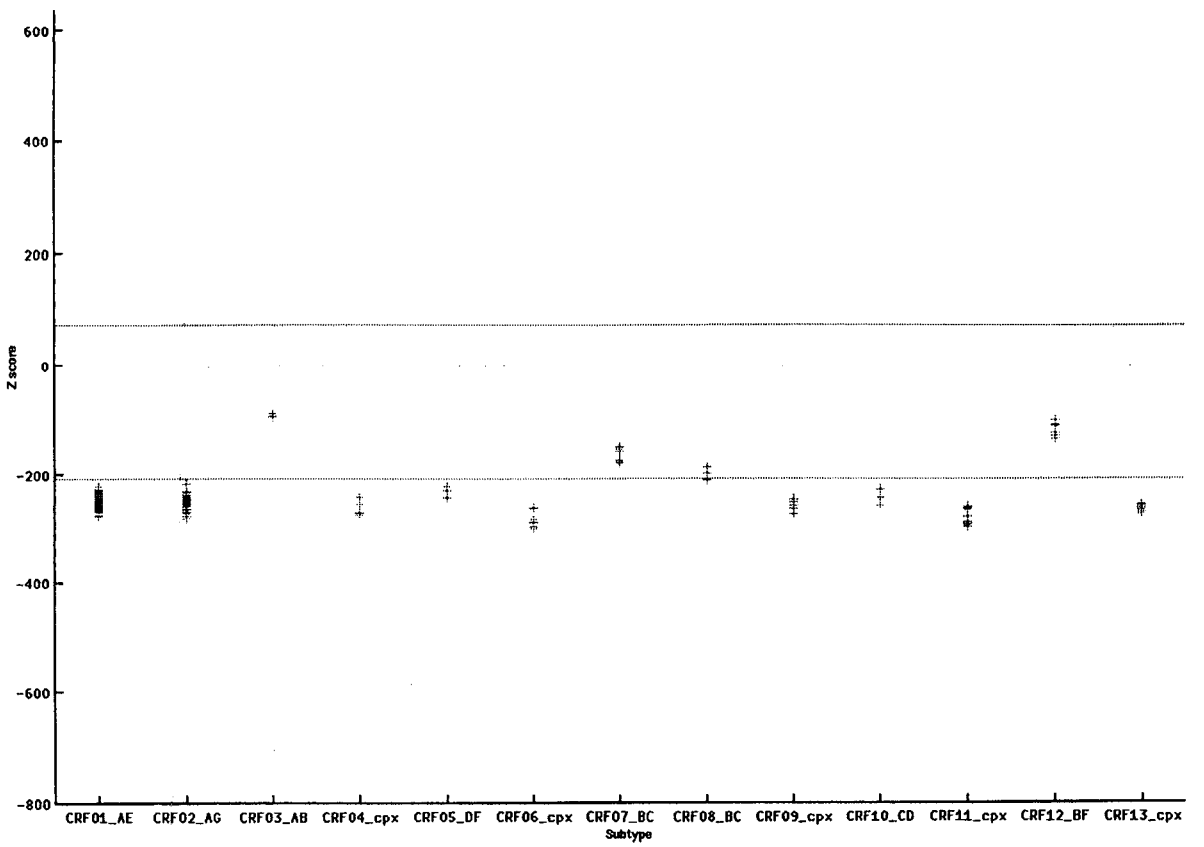


Figure (4.2)

In this plot, level of different types of CRF strains from CRF14 to CRF26 are shown to detect those strains which are recombination of some parts of gag-pol segment B subtype. The strains belonging to type CRF17_BF return Z score between Tp and Tn. This prediction under our assumption shows that B subtype is present in its composition. Each strain of the types CRF20_BG, CRF23_BG and CRF24_BG return Z score either above or just near the threshold Tn which indicate that this type of strains have made of minimal length segment of B subtype. The strains belonging to each type CRF14_BG and CRF_01B does not return Z score above the threshold Tn therefore according to our assumption, B subtype strains not present in its gag-pol coding region .

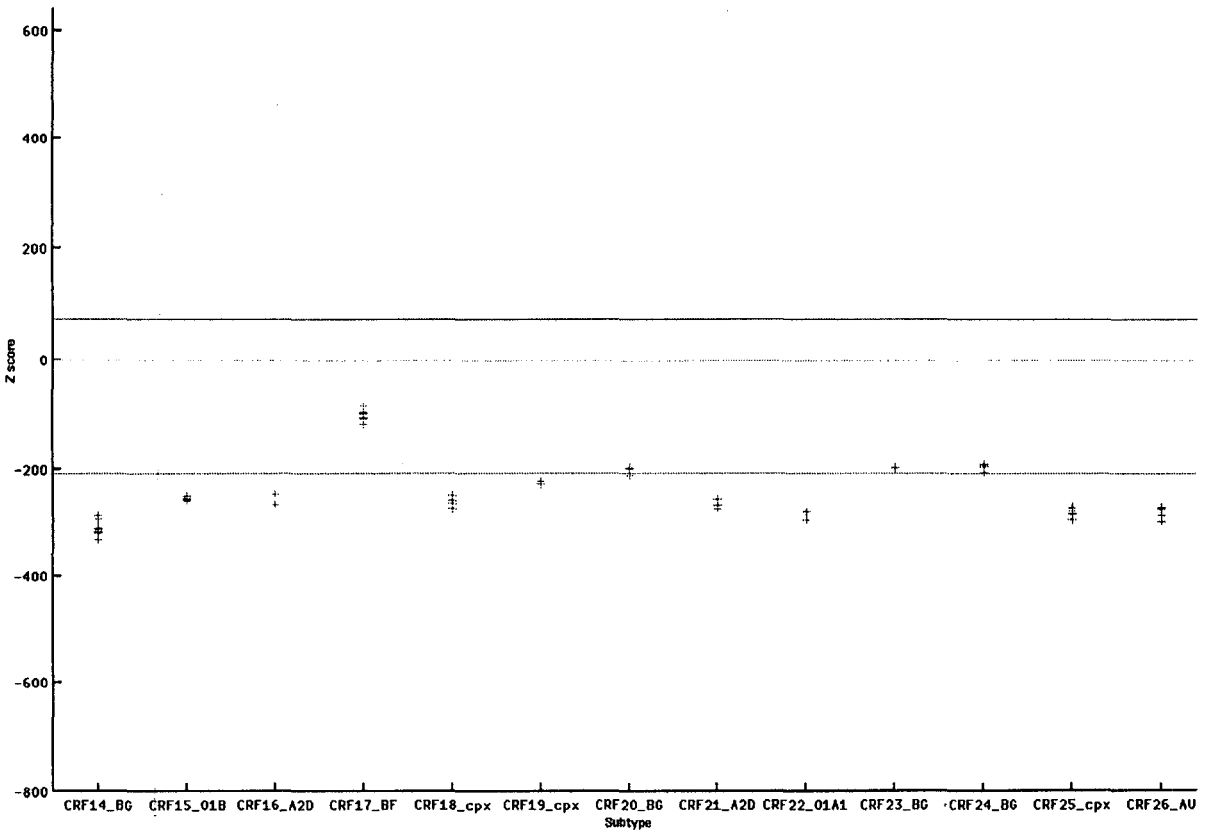


Figure (4.3)

In this plot level of different types of CRF strains from CRF27 to CRF39 are shown to detect those strains which contain some length of segment B subtype. All the strains belonging to CRF39_BF return higher than Tp, since gag-pol coding region of CRF39_BF is primarily made up of sections from the gag-pol region of B subtype and has very short lengths of gag-pol coding region of F subtype. All the strains of CRF28_BF return Z score near the threshold Tp and also in the case, when method is used to detect F subtype containing CRF, these type strains return Z score between corresponding thresholds Tp and Tn. Therefore these strains have maximal length and minimal length of segment of gag-pol coding region of B subtype and F subtype respectively. The strains belongs to CRF29_BF, CRF33_01B and CRF34_01B return Z score between Tp and Tn, which allows us to predict that B subtype is part of its composition. The strains belonging to type CRF31_BC return Z score near Tn which indicate that this type of strains have made of very short length segment of gag-pol coding region of B subtype.

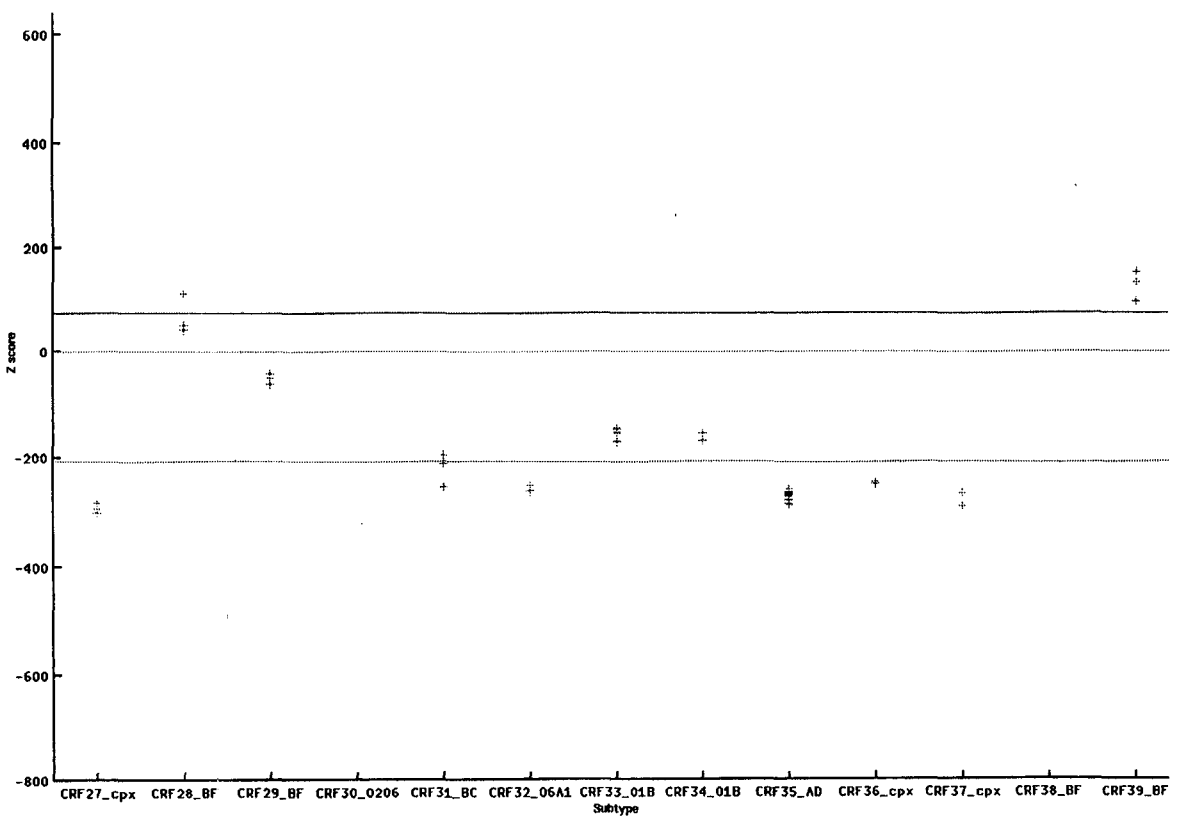


Figure (4.4)

In this plot, level of different types CRF strains from CRF40 to CRF48 are shown to detect those strains which are recombination as a fraction of gag-pol segment of B subtype. The strains belonging to types CRF42_BF and CRF47_BF return Z scores between Tp and Tn as well as just near the threshold (Z=0) which shows that maximal length of gag-pol segment of B subtype involve to recombine in their gag-pol coding region. The CRF40_BF and CRF44_BF return Z score between Tp and Tn, this prediction according to the our assumption shows that B subtype involve to recombine as a part in its composition.

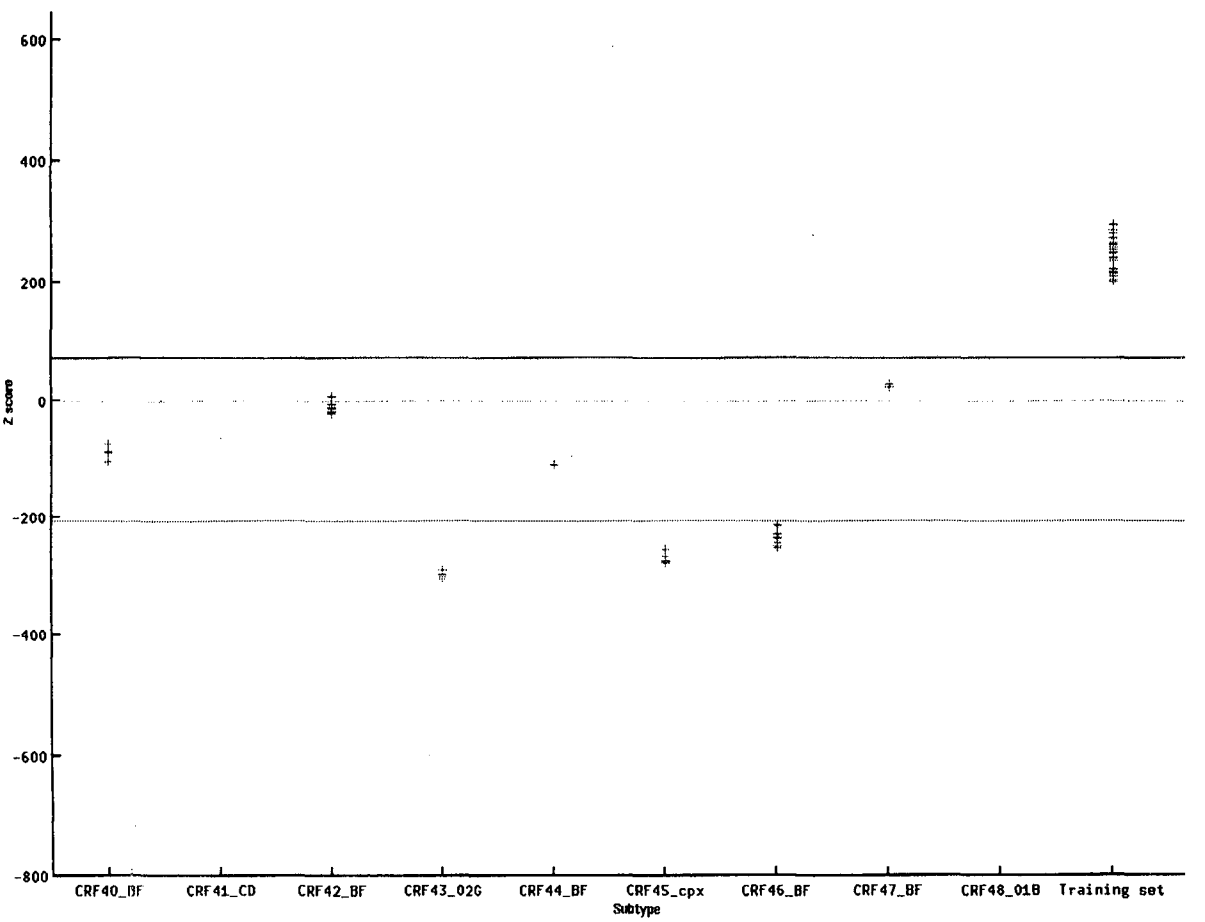


Figure (4.5)

4.3 Result for detection of CRF strains which contains F subtype strains:

In this section, we have shown that detection of each query sequences belonging to the test set which make up as a part of F subtype in its composition. There are five plots to present the distribution of Z scores, which generated for detection of F subtype by the extended method using pHMMs, with different pure subtypes as well as different types of CRF strains. In this case, we have taken eight sequences of gag-pol coding region of F1 sub-subtype and two sequences of gag-pol coding region of F2 subsubtype in positive training set. We take two sequences of each subtype except F1 and F2 sub-subtype in negative training set of F subtype.

Distribution of Z scores and two thresholds which are Tp and Tn are demarcated in following graph, these are defined in the chapter "Method". This plot shows all the maximum Z scores of each different subtype are nearly same which define the threshold Tn for F subtype.

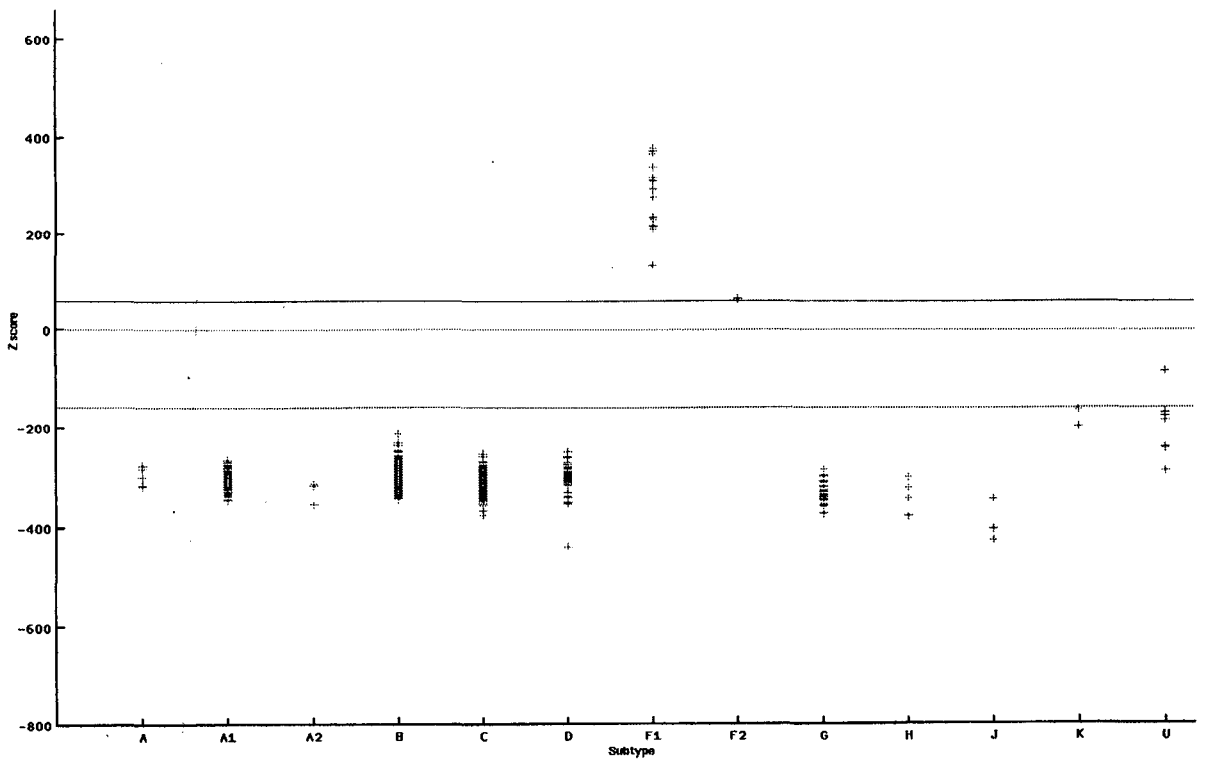


Figure (4.6)

In this plot, level of different types of CRF strains from CRF01 to CRF13 are shown to detect those strains which are recombination as part of F subtype. All the strains belonging to type CRF12_BF return higher than Tp therefore according to our assumption these are pure strains of F subtype. Therefore we conclude that these type of strains are made up with B subtype and F subtype in which gag-pol segment of F subtype is maximal. Each the subtype of CRF05_DF and are made up from F subtype .

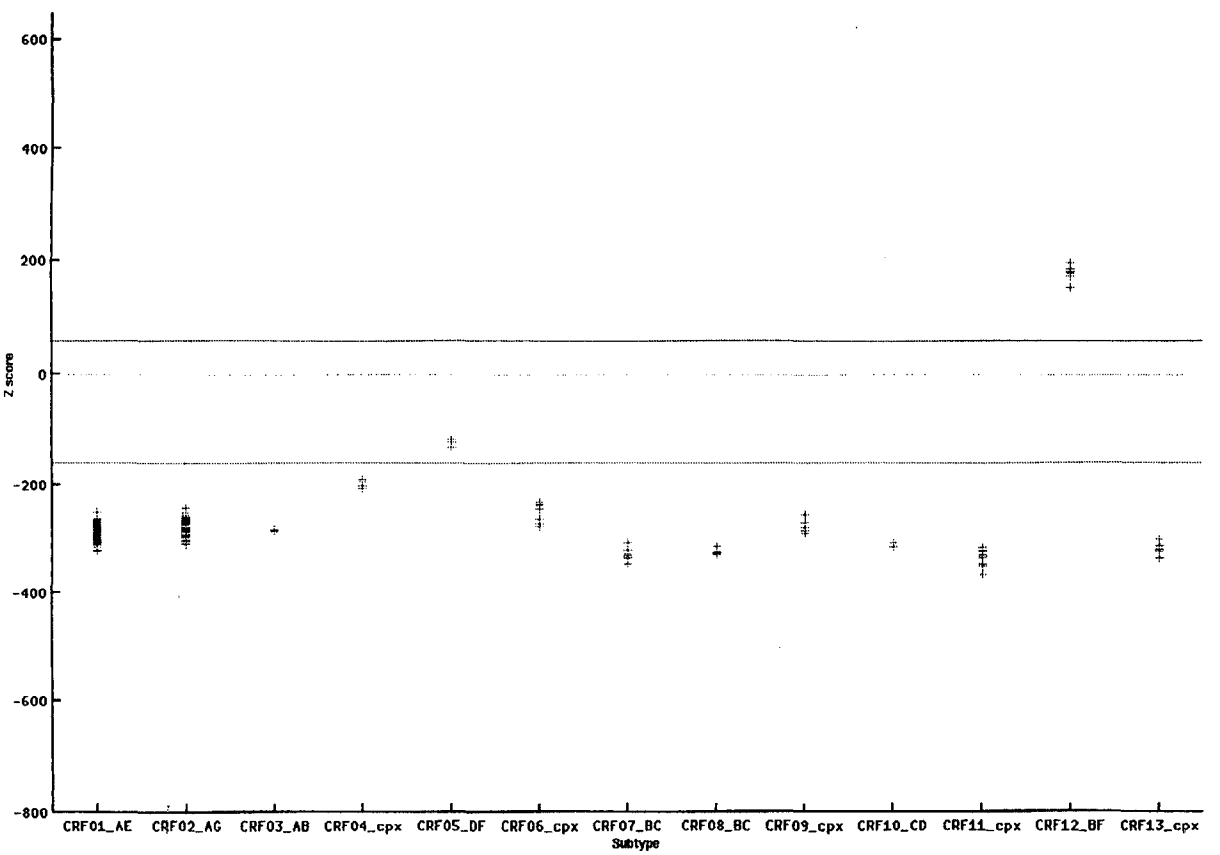


Figure (4.7)

In this plot, level of different types of CRF strains from CRF14 to CRF26 are shown to detect those strains which are recombination as a fraction of gag-pol segment F subtype. All the strains belonging to CRF17_BF return higher than Tp therefore according to our assumption, these are pure strains of F subtype. But we see earlier in results of detection of B subtype, these strains are made up as a fraction of gag-pol segment of B subtype. Therefore we conclude that these type of strains are made up with B subtype and F subtype in which length of gag-pol segment of F subtype is maximal.

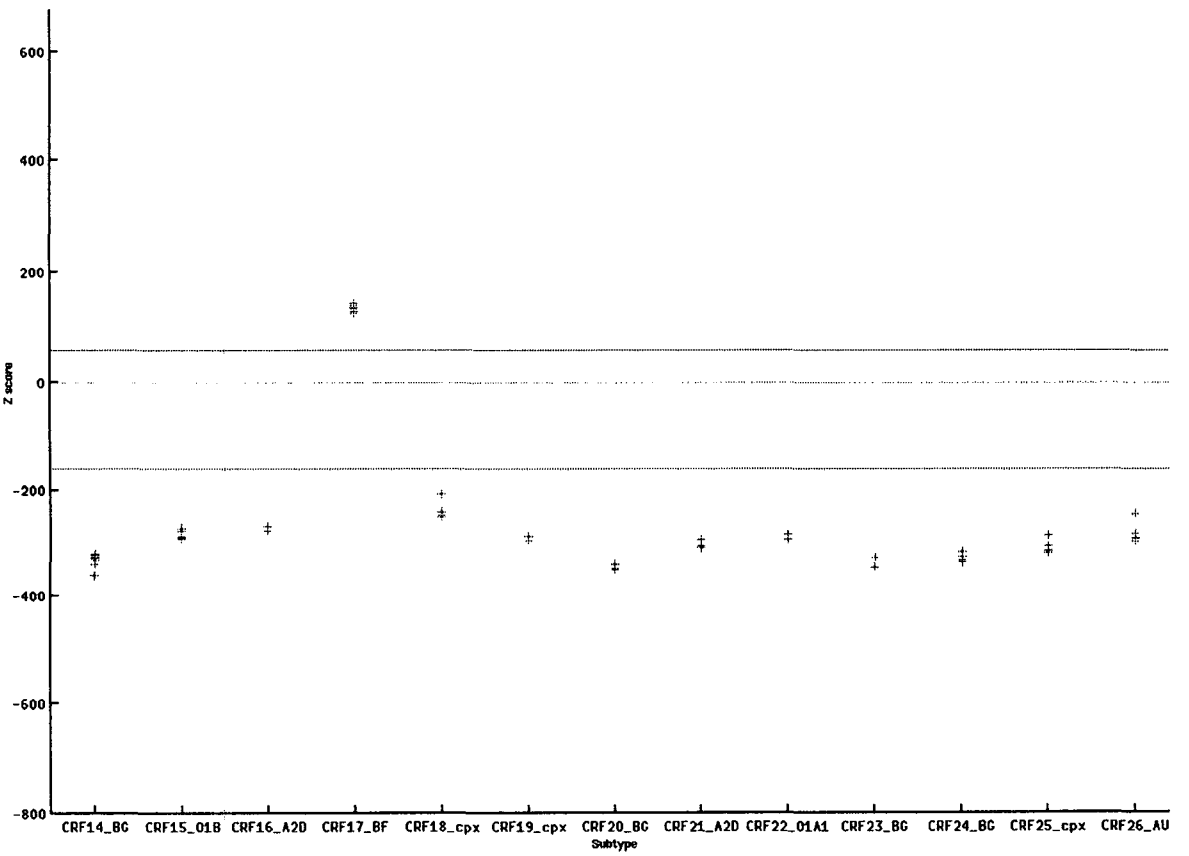


Figure (4.8)

In this plot, level of different types of CRF strains from CRF27 to CRF39 are shown to detect those strains which are recombination as a part of some length of gag-pol segment F subtype. All the strains belonging to CRF28_BF and CRF29_BF return higher than threshold (Z=0) and also some members of it return Z score higher than Tp if we conclude, to combine the results of detection of B subtype for these type of strains, that these strains are mixture of F subtype and B subtype. Problem is arisen in detection of F subtype in CRF39_BF because the total lengths of different segments of F subtype is very small in gag-pol coding region.

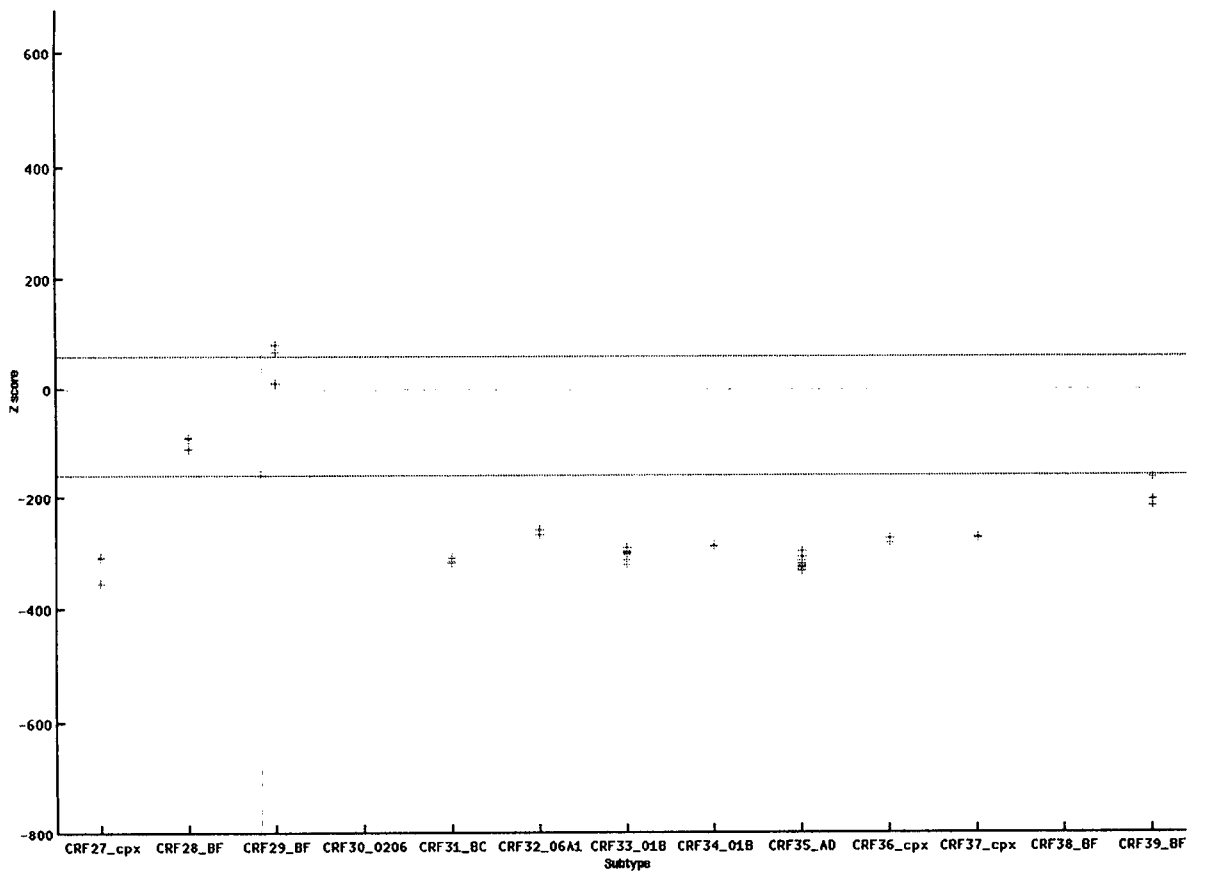


Figure (4.9)

In this plot, level of different types of CRF strains from CRF40 to CRF48 are shown to detect those strains which are recombination of some part of gag-pol segment of B subtype. All the strains belonging to CRF44_BF and CRF46_BF return higher than Tp. Therefore we conclude that these type of strains are made up with B subtype and F subtype in which length of gag-pol segment of F subtype is maximal. The CRF40_BF, CRF42_BF and CRF47_BF return Z score between Tp and Tn, this prediction under the our assumption shows that F subtype involve to recombine as a part in its composition.

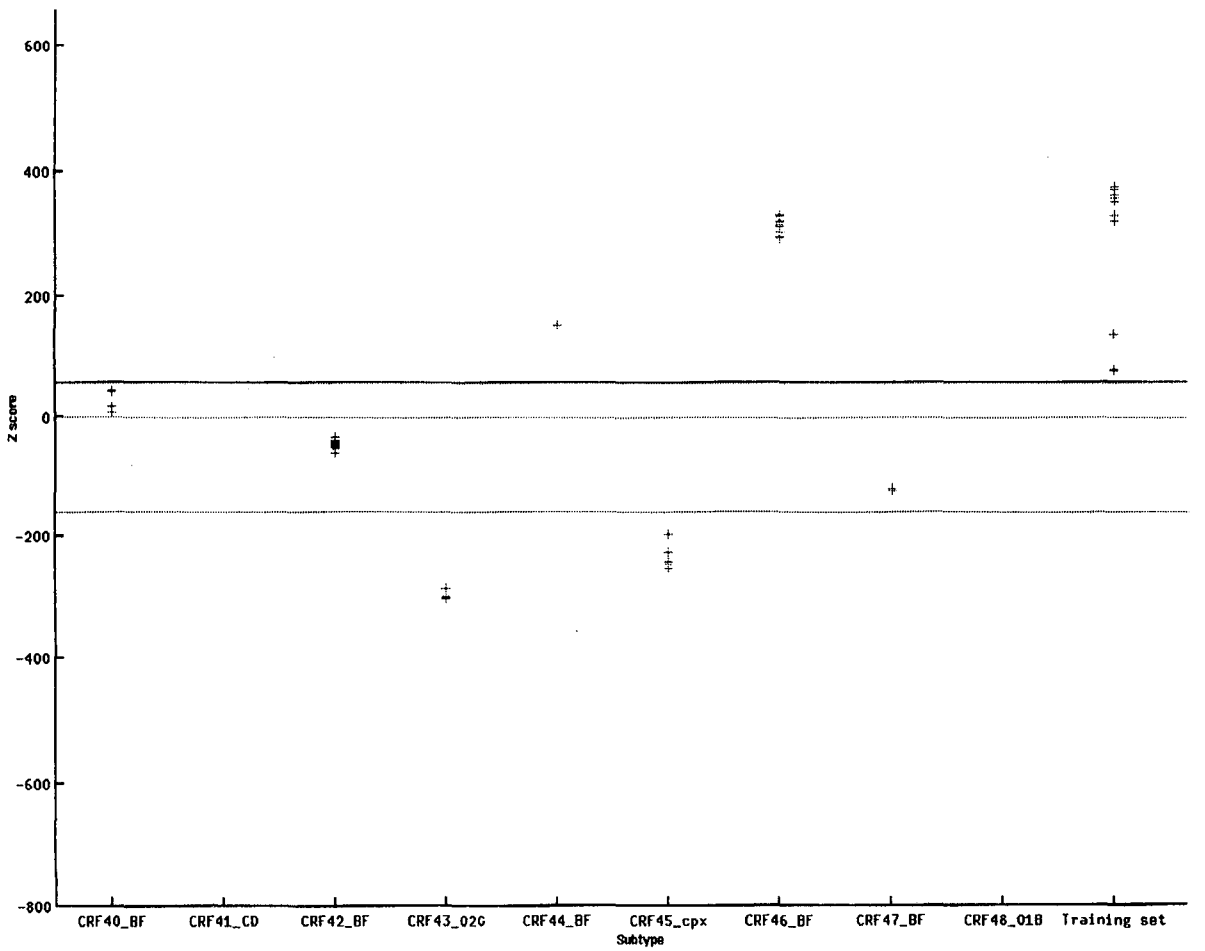


Fig (4.10)

4.4 The table of thresholds for detection of CRF strains:

Subtype	Threshold (Tp)	Threshold (Tn)
A	177	-220
B	74	-207
C	101	-300
D	74	-140
F	60	-160
G	170	-325

Chapter 5

Discussion and Conclusion

We have shown in previous chapters, how the pHMM method utilizing both negative and positive training set using pHMMs can lead to substantial improvements in the accuracy of detection of different subtypes. We also saw that the number of sequences of different subtypes of negative training set have played important role in the identification of different subtypes that make up a CRF strain. In this chapter we briefly explore performance and advantage of this method which allows detection of different strains of HIV-1 with higher degree of reliability.

Initially, we used the standard method of classification using pHMMs. Even though this method performed well for classification of most subtypes, the accuracy of classification was substantially low for B and D subtypes which are more closely related to each other in comparison to any two different subtypes. Another disadvantage of this method is that the thresholds for each subtype are not fixed as a result of which the robustness of the classifier can change depending on the threshold selected.

We then used an efficient, reliable and robust method using pHMMs which generate Z score of a query sequence to determine its subtype on the basis of similarity of gag-pol coding region. This method works effectively on classification of strains of different subtypes including more similar subtypes ,except D subtype , of HIV-1. For accurate detection of D subtype, we changed the number of B subtype sequences form two to six in its negative training set. The Z score of query sequence for

given subtype classifier (pHMM) coupled with fixed threshold (Z=0) is sufficient to predict whether or not the query sequence is a member of that subtype. either assigned to associated subtype or is not assigned to that subtype. Although we used the gag-pol segment to build our pHMMs, this method can be easily adapted for classification on the basis of the env protein coding region.

The pHMMs also provide an efficient method for detecting the subtype composition of CRF strains. The detection is based on the Z score relative to the thresholds Tp and Tn. Typically, a query CRF sequence with a Z-score (estimated for a particular subtype classifier) lying between Tn and Tp would be considered to be contain segment of the corresponding subtype in its sequence. Even though we used this method to detect recombination events that occurred in the gag-pol region, the method can be easily extended to detect recombination events occurring outside of the gag-pol region by choosing different protein-coding regions of the HIV genome to build our pHMM models. The pHMM based methods described in this thesis does not require the accurate construction of reliable phylogenetic trees. Another advantage is that it requires only a small number of training sequences to return classification results with high accuracy. Hence it can be effectively used in accurately classifying subtypes for which only a small number of sequences are available profile HMM construction.

5.1 Ongoing and future work:

This is general method which can be applied for classification of intra-species variation of other organisms. We successfully classified the different genotypes of HBV viruses on the basis of complete genome variation. All the genotypes except H of HBV were successfully detected on the basis of Z score when coupled with threshold (Z=0). A problem arose in detection of strains of H genotype due to similarity of the strains of H subtype and strains of F genotype. The problem was eliminated by taking the six genome sequence of F subtype and two genome sequences of all the genotype except for H and F which are in negative training set of H genotype. We are currently continuing our analysis of the HBV classification using this approach.

References

[1] **Joint United Nations Program on HIV/AIDS (UNAIDS) and World Health Organization (WHO):** AIDS epidemic update. 2009.

[2] **Tebit DM, Nankya I, Arts EJ and Gao Y:** HIV diversity,recombination and disease progression: how does fitness "fit" into the puzzle?. AIDS Rev 2007, 9:75–87.

[3] **Preston B, Poiesz B and Loeb L:** Fidelity of HIV-1 reverse transcriptase. Science 1988, 242:1168–1171.

[4] **Domingo E and Holland J:** RNA virus mutations and fitness for survival. Annu Rev Microbiol 1997, 51:151–178.

[5] **Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B,Funkhouser RK, Gao F, Hahn BH, Kuiken C and Learn GH, et al:** HIV-1 Nomenclature Proposal. Human Retroviruses and AIDS 1999 New Mexico: Los Alamos National Laboratory: **Kuiken CL, Foley B, Hahn B, Korber B, McCutchan F, Marx PA, Mellors JW, Mullins JI, Sodroski J, Wolinksy S** 1999, 492–505.

[6] **Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C and Detours V:** Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull 2001, 58: 19–42.

[7] **Hutchinson JF:** The biology and evolution of HIV. Annu. Rev. Anthropol. 2001, 30: 85–108.

[8] **Myers G, MacInnes K and Korber B:** The emergence of simian/human immunodeficiency viruses. AIDS Res Hum Retroviruses 1992, 8:373–386.

[9] **Louwagie J, McCutchan FE, Peeters M, Brennan TP, Sanders-Buell E, Eddy GA, Groen vander G, Fransen K, Gershy-Damet GM and Deleys R, et al:** Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. AIDS 1993, 7:769–780.

[10] **Janssens W, Heyndrickx L, Fransen K, Motte J, Peeters M, Nkengasong JN, Ndumbe PM, Delaporte E, Perret JL and Atende C, et al:** Genetic and phylogenetic analysis of env subtypes G and H in central Africa. AIDS Res Hum Retroviruses1994, 10:877–879.

[11] **Triques K, Bourgeois A, Vidal N, Mpoudi-Ngole E, Mulanga-Kabeya C, Nzilambi N, Torimiro N, Saman E, Delaporte E and Peeters M:** Near-full-length genome sequencing of divergent African HIV-1 subtype F viruses leads to the identification of a new HIV-1 subtype designated K. AIDS Res Hum Retroviruses 2000, 16:139–151.

[12] **Gao F, Robertson DL, Carruthers CD, Li Y, Bailes E, Kostrikis LG, Salminen MO, Bibollet-Ruche F, Peeters M and Ho DD, et al:** An isolate of human immunodeficiency virus type 1 originally classified as subtype I represents a complex mosaic comprising three different group M subtypes (A, G, and I).Publish with Bio Med Central and every J Virol 1998, 72:10234–10241.

[13] ***Triques K, A Bourgeois, N Vidal, E Mpoudi-Ngole, C Mulanga-Kabeya, N Nzilambi, N Torimiro, E Saman, E Delaporte, and M Peeters:*** Near-full length genome sequencing of divergent African HIV-1 subtype F viruses leads to the identification of a new HIV-1 subtype designated K. ARHR 2000, 16:139-151.

[14] **Myers RE, Gale CV, Harrison A, Takeuchi Y and Kellam P:** A statistical model for HIV-1 sequence classification using the subtype analyser (STAR). Bioinformatics 2005, 21:3535–3540.

[15] **Krogh A, Brown M, Mian I, Sjolander K, Haussler D:** Hidden Markov Models in Computational Biology: Applications to protein modelling. J Mol Biology235:1501-1531 1994.

[16] **Rabiner LR:** A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77, 1989, 257–286.

[17] **Durbin R, Eddy SR, Krogh A and Mitchison GJ:**Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (Cambridge University Press, Cambridge UK), 1998

[18] **Eddy SR:** HMMER: Profile Hidden Markov Models for biological sequence analysis. 2001. **HMMER 3.0** (28 March 2010) [http://hmmer.janelia.org/]

[19] **Edgar RC:** MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004, 32, 1792–1797

Appendix

The code (written in perl) to calculate Z scores of different types of HIV-1 strains :

```
#!/usr/bin/perl

print "please enter the name of positive file\t";
$first=<>; # Enter the  name of file containing scores generated by positive  pHMM
print "please enter the name of negative file\t";
$second=<>; # Enter the  name of file containing scores generated by negative pHMM
open FF, "$first" or die $!;
@ar1=<FF>;
close(FF);
open SS, "$second" or die $!;
@ar2=<SS>;
close(SS);
print "please enter the name of file conaining positive training set \t";
$tr=<>; # Enter the positive training set
open TR, "$tr" or die $! ;
@tra=<TR>;
close(TR);
$c=0;
foreach$trai(@tra)     # extraction of sequence I.D. from positive training set
{
```

```
if(substr($trai,0,1) eq '>')
        {
                $train[$c]= substr($trai,1,length($trai));
                $c++;
        }
}
print "please enter the name of output file \t " ;
$out=<>; # Enter the file name which to store the Z scores with its associated label
open XX, ">$out" or die $!;
########################################
#                                      #
# Arrangement I.D.of sequences of positive  #
# and negative trainig set which are        #
# broken by HMMER 3                         #
#                                       #
########################################

for($i=0;$i< @ar1;$i++)
{
@a1=split(' ',$ar1[$i]);@a2 = split(' ',$ar2[$i]);
for ($k=@a1-1;$k>7;$k--)
{
 $st =$a1[$k].$st; $st1=$a2[$k].$st1;
}
for ($k=0;$k<7;$k++)
        {
          $array1[$k]= $a1[$k] ;
        $array2[$k]= $a2[$k];

         } $k=0;
```

```
        $array1[7]=$st; $array2[7]= $st1;
        $st='';$st1='';
    for ($k1=0;$k1<=7;$k1++)
        {
            $A[$i][$k1]=$array1[$k1];
            $A1[$i][$k1]=$array2[$k1];
        }
}
############################################################
#                                                          #
# Training sequences are assigned to number 6200 with      #
# corresponding Z score for its demarcation                #
#                                                          #
############################################################

for($u=0;$u<$c;$u++)
{ chomp($train[$u]);
        for($q=0;$q < @ar1;$q++)
          {
         for($q1=0;$q1< @ar2;$q1++)
                {
                    if($A[$q][7] eq $train[$u] && $A1[$q1][7] eq $train[$u])
                     {    $A[$q][7]= '';
                          $di1= $A[$q][1]-$A1[$q][1];
                          $vars=6200;
                          print XX "$di1\t$vars\n" ;
                          $A[$q][7]='';
                      }
                    }
          }
```

```
}
        ####################################
        #                                  #
        # Different types of pure and CRF strains  #
        # are assigned too different corresponding  #
        # numbers with their associated Z scores   #
        #                                  #
        ####################################
for($l=0;$l<@ar1;$l++)
 {  for ($j=0;$j<@ar2;$j++)
        {
                if($A[$l][7] eq $A1[$j][7] )
                {
                        @v=split(",$A[$l][7]);
                        $di= $A[$l][1]-$A1[$j][1];
                                if($v[0] eq'B' )
                                        {       $vars=400;
                                                print XX "$di\t$vars\n" ;
                                         }
                                                elsif($v[0] eq'A' && $v[1] eq '.' )
                                                {       $vars=100;
                                                        print XX "$di\t$vars\n" ;
                                                }
                                        elsif($v[0] eq'A' && $v[1] eq '1' )
                                        {       $vars=200;
                                                print XX "$di\t$vars\n" ;
                                        }
                                        elsif($v[0] eq'A' && $v[1] eq'2' )
                                        {
                                                $vars=300;
```

```
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'C' && $v[1]eq'.')
{       $vars=500;
        print XX "$di\t$vars\n" ;
}

elsif($v[0] eq'D' && $v[1]eq'.')
{       $vars=600;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'F' && $v[1]eq'1')
{       $vars=700;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'F' && $v[1]eq'2')
{       $vars=800;
        print XX "$di\t$vars\n" ;
}

elsif($v[0] eq'H' && $v[1]eq'.')
{       $vars=1000;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'G' && $v[1]eq'.')
{       $vars=900;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'K' && $v[1]eq'.')
{
```

```
        $vars=1200;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'J' && $v[1]eq'.')
{       $vars=1100;
        print XX "$di\t$vars\n" ;
}


elsif($v[0] eq'U' && $v[1]eq'.')
{       $vars=1300;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'0' && $v[1]eq'1')
{       $vars=1400;
        print XX "$di\t$vars\n" ;
}

elsif($v[0] eq'0' && $v[1]eq'2')
{       $vars=1500;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'0' && $v[1]eq'3')
{       $vars=1600;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'0' && $v[1]eq'4')
{
        $vars=1700;
        print XX "$di\t$vars\n" ;
```

```
}
elsif($v[0] eq'0' && $v[1]eq'5')
{       $vars=1800;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'0' && $v[1]eq'6')
{       $vars=1900;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'0' && $v[1]eq'7')
{       $vars=2000;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'0' && $v[1]eq'8')
{       $vars=2100;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'0' && $v[1]eq'9')
{       $vars=2200;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'1' && $v[1]eq'0')
{       $vars=2300;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'1' && $v[1]eq'1')
{       $vars=2400;
        print XX "$di\t$vars\n" ;
}
```

```
elsif($v[0] eq'1' && $v[1]eq'2')
{       $vars=2500;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'1' && $v[1]eq'3')
{       $vars=2600;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'1' && $v[1]eq'4')
{
        $vars=2700;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'1' && $v[1]eq'5')
{       $vars=2800;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'1' && $v[1]eq'6')
{       $vars=2900;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'1' && $v[1]eq'7')
{       $vars=3000;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'1' && $v[1]eq'8')
{       $vars=3100;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'1' && $v[1]eq'9')
```

```
{       $vars=3200;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'2' && $v[1]eq'0')
{       $vars=3300;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'2' && $v[1]eq'1')
{       $vars=3400;
        print XX "$di\t$vars\n" ;
}

elsif($v[0] eq'2' && $v[1]eq'2')
{       $vars=3500;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'2' && $v[1]eq'3')
{       $vars=3600;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'2' && $v[1]eq'4')
{
        $vars=3700;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'2' && $v[1]eq'5')
{       $vars=3800;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'2' && $v[1]eq'6')
```

```
{       $vars=3900;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'2' && $v[1]eq'7')
{       $vars=4000;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'2' && $v[1]eq'8')
{       $vars=4100;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'2' && $v[1]eq'9')
{       $vars=4200;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'3' && $v[1]eq'0')
{       $vars=4300;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'3' && $v[1]eq'1')
{       $vars=4400;
        print XX "$di\t$vars\n" ;
}

elsif($v[0] eq'3' && $v[1]eq'2')
{       $vars=4500;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'3' && $v[1]eq'3')
{       $vars=4600;
```

```
print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'3' && $v[1]eq'4')
{
        $vars=4700;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'3' && $v[1]eq'5')
{       $vars=4800;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'3' && $v[1]eq'6')
{       $vars=4900;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'3' && $v[1]eq'7')
{       $vars=5000;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'3' && $v[1]eq'8')
{       $vars=5100;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'3' && $v[1]eq'9')
{       $vars=5200;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'4' && $v[1]eq'0')
{       $vars=5300;
        print XX "$di\t$vars\n" ;
```

```
}
elsif($v[0] eq'4' && $v[1]eq'1')
{       $vars=5400;
        print XX "$di\t$vars\n" ;
}

elsif($v[0] eq'4' && $v[1]eq'2')
{       $vars=5500;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'4' && $v[1]eq'3')
{       $vars=5600;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'4' && $v[1]eq'4')
{
        $vars=5700;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'4' && $v[1]eq'5')
{       $vars=5800;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'4' && $v[1]eq'6')
{       $vars=5900;
        print XX "$di\t$vars\n" ;
}
elsif($v[0] eq'4' && $v[1]eq'7')
{       $vars=6000;
        print XX "$di\t$vars\n" ;
```

```
                    }
                    elsif($v[0] eq'4' && $v[1]eq'8')
                    {       $vars=6100;
                            print XX "$di\t$vars\n" ;
                    }
            }
    }
}
            close(XX)       ;
```