Prediction of conserved peptides containing multiple T and B cell epitopes in Ebola virus glycoprotein

Dissertation

Submitted in the partial fulfillment of the requirement for

the award of the degree of

MASTER OF TECHNOLOGY

IN

BIOTECHNOLOGY

By

Shinjan Kapur

601304014

Under the guidance of

Dr. Manoj Baranwal

Assistant Professor

DEPARTMENT OF BIOTECHNOLOGY

THAPAR UNIVERSITY, PATIALA (PUNJAB)-147004

JULY 2015

CERTIFICATE

This is to certify that the project entitled "**Prediction of conserved peptides containing multiple T and B cell epitopes in Ebola virus glycoprotein**" being submitted by Shinjan Kapur, in partial fulfillment of the requirements for the award of degree of Master of Technology in Biotechnology to Thapar University, Patiala, is a record of student's own work carried out by her. The report has not been submitted for the award of any other degree or certificate in this or any other university.

Dr Manoj Baranwal
Assistant Professor, Supervisor
Department of Biotechnology
Thapar University, Patiala

Dr. Dinesh Goyal
Professor & Head
Department of Biotechnology
Thapar University, Patiala

Dr. S.S Bhatia
Dean
Academic Affairs,
Thapar University
Patiala

CANDIDATE'S DECLARATION

I, hereby declare that the work presented in the thesis entitled **"**In silico Prediction of conserved immunogenic peptides containing overlapping $CD4^+$ and $CD8^+$ T cell epitopes in Ebola virus envelope glycoprotein (GP)**"** in the partial fulfillment of the requirement for the award of the degree of Master of Technology in Biotechnology, Department of Biotechnology, Thapar University, Patiala, is an authentic record of my work during the period of one year from July 2014 to June 2015, under the guidance of Dr. Manoj Baranwal, Assistant Professor, Thapar University, Patiala. I have not submitted the matter embodied in this thesis for the award of any other degree or diploma.

Place: Patiala
Date: 14/07/15

Shinjan Kapur
(601304014)

ACKNOWLEDGEMENT

This is a golden opportunity for me to convey my sincere regards to all those people who enabled me to accomplish my project successfully.

It is my privilege to express my sincere gratitude to my supervisor, **Dr. Manoj Baranwal**, Assistant Professor, for his encouragement and valuable advice. I am thankful that he gave me a brilliant opportunity to conduct my M.Tech thesis under his guidance. His confidence imbibing attitude, continuous support, valuable advice and scientific approach helped me to a great extend to accomplish my work.

I express my sincere thanks to **Dr. Dinesh Goyal**, Professor and Head, Department of Biotechnology, Thapar University, Patiala, for providing me with the best laboratory facilities during my research pursuit.

Special thanks to **Miss. Neha Lohia** research scholar (DBT), for her immense help and valuable suggestions throughout my study.

I also express my sincere thanks to **Vivek Sharma, Sukhjeet Kaur** and **Vikrant Mehta** for their prompt suggestions and support throughout the Project. I thank profusely all the laboratory staff, especially **Mr**. **Ram Newal** and Mr. **Lallan** for their kind help and co-operation throughout study.

Last but not the least, I am thankful to my friends **Maitri Bhandari** and **Vikas Batra**, for their constant encouragement as it motivated me to work hard and give my best efforts during my research project.

Last but not the least I am grateful to my **family** for their love, concern guidance and support throughout my project.

Shinjan Kapur

601304014

Table of Contents

A

ABSTRACT

Glycoprotein (GP) is the surface protein of Ebola virus (EBOV), responsible for penetration of the virus into the host cell cytoplasm by mediating the fusion of the membrane of the endocytosed virus particle with the endosomal membrane. Hence EBOV GP is considered to be the main target for design of vaccine and entry inhibitors. Due to extreme pathogenicity of EBOV, a high degree of biohazard containment is required for laboratory studies and clinical analysis and thus the conventional approach of vaccine and therapeutic development is hampered. Thus a cost and time effective approach needs to be adopted. Present study uses an immunoinformatics approach to identify conserved peptides containing multiple T and B cell epitopes of GP protein in three virulent species ***Zaire ebolavirus, Sudan ebolavirus and Bundibugyo ebolavirus*** using sequences available till August 2014. Different prediction algorithms have been used for peptide containing multiple epitopes and four conserved peptides were identified containing multiple epitopes for both T and B cells. The fragment **LRQLANETTQALQLFLRATTELRTFSILNRKAIDFLL**, out of the final four peptide fragments, contained maximum number of predicted epitopes for CD8+ and CD4+ cells. After studying the crystal structure of GP protein, it was found that this peptide lies in the heptad repeat regions HR1 and HR2 of GP2. Population coverage analysis has shown that all four peptides have the capacity to induce immune response among the different populations of the world and they are found to have binding affinity to large number of HLA alleles. Molecular docking studies revealed that all four peptides have comparable binding energy with that of the native peptide. Thus these peptides can further be evaluated for their potency as vaccine development against Ebola hemorrhagic fever.

B

LIST OF ABBREVIATIONS

FDA	Food and Drug Administration
CDC	Centers for Disease Prevention and Control
WHO	World Health Organization
ZEBOV	Zaire Ebola Virus
SUDV	Sudan Ebola Virus
TAFV	Tai Forest Ebola virus
BDBV	Bundibugyo Ebola Virus
RESTV	Reston Ebola Virus
GP	Glycoprotein
MHC	Major Histocompatibility Complex
QSAR	Quantitative Structure Activity Relationship
KDa	Kilo Daltons

C

LIST OF TABLES

D

LIST OF FIGURES

Chapter 1 INTRODUCTION

Ebola Virus is a prototype pathogen of an acute and fatal type of hemorrhagic fever in humans designated as the Ebola viral disease. It was first identified in a village near Ebola River in 1976 from which the disease take its name. The 2014 West Africa Ebola outbreak is the largest Ebola outbreak with countries like Liberia, Guinea and Sierra Leone being most severely affected. This outbreak is also the first Ebola epidemic in the world with about 11,232 confirmed deaths as on 26th June 2015 (Center of disease control and prevention). Natural host of Ebola is still unknown though a prominent suspect is fruit bats which are consumed in soup in West Africa. With high fatality rate and absence of treatment and vaccination options, Ebola virus is declared as a public health pathogen and biothreat pathogen of Category A (Feldmann *et al*., 2011)

Although the clinical course of infection is known but the complete molecular biology and pathogenesis of Ebola virus virulence factors and host cells responses is still a factor of ongoing research. This is due to various factors like, disease outbreaks occurring in remote areas where the collection of samples is difficult. Also the samples of Ebola are highly contagious which demands a very strict sterilization protocol while collection and a high degree of biohazard containment for studies in laboratories and clinical analysis

Ebola Virus belongs to the Filoviridae family. Filovirus are among the most lethal human pathogens known till date with fatality rates ranging up to 90% depending upon the virus species and strain. Filovirus consist of two genera: *Ebola virus* and *Marburg Virus* (Changula *et al.*, 2013). These are non segmented, enveloped RNA viruses with negative strand and characteristic filamentous particle (Feldmann et al., 2011) The genus *Ebolavirus* consists of five identified species, *Zaire Ebolavirus*, which is associated with 2014 Ebola outbreak and with highest fatality rates, *Sudan Ebolavirus*, *Bundibugayo Ebolavirus*, *Taï forest ebolavirus* and *Reston ebolavirus*. There are no reported cases of human infection by *Reston Ebolavirus*. The Ebola virus genome is 19Kb long and consists of seven genes: Nucleoprotein (NP), which is the most abundant protein in the viral cell,

Glycoprotein (GP) which is the only surface protein, VP35 VP40, VP30, VP24 and polymerase genes (Changula *et al.,* 2013) GP gene after transcriptional editing expresses two gene products: a 676- residues transmembrane linked glycoprotein (GP) and a 364-residues secreted glycoprotein (sGP). GP is cleaved post-translationally by Furin into GP_1 and GP_2 which are disulphide linked. Trimer of $GP_{1,2}$ form spikes present on viral surface. GP_1 is responsible for cell surface attachment to host cells whereas GP_2 mediated fusion with host membranes. Several evidences suggest that Glycoprotein play a key role in manifestation of the disease. While sGP is suspected to alter immune response by inhibiting the activation of Neutrophils, GP contribute in the manifestations of Hemorrhage symptoms (Sullivan, *et al.,* 2011) Hence EBOV GP is an ideal potential target for designing vaccines and entry inhibitors (Jeffrey E. Lee *et al.,* 2008)

During EBOV infection, virus encoded glycoprotein are released in the host cells in soluble form. High levels of truncated surface GP and sGP can be detected in the blood of infected patients. Primary infection starts with the replication of the virus in the dendritic cells, macrophages and monocytes followed by replication in hepatocytes and splenocytes. This is followed by massive release of cytokines, chemokines and vasoactive substances leading to inflammation disorders (Escudero-Pérez B *et al.,* 2014). Glycoprotein binds to the endothelial cells leading to cytokine dysregulation and loss of vascular integrity leading to endothelial leakage and vasomotor collapse leading to rapid death of the host (Sullivan *et al.,* 2003). Extremely lethal nature of this virus hampers the study of live virus and requires extremely high sterilization conditions. There is no licensed vaccine, proven treatment and rapid diagnostic available for Ebola till date. Thus therapeutics identification is of utmost priority. Two experimental vaccines have entered human clinical trials: ChAd3-ZEBOV and rVSV-ZEBOV.

It has been reported that $CD8^+$ cells play a significant role in initial clearance of virus during acute infection while B-cells and $CD4^+$ cells are required for long term protection and for avoiding the reemergence of the disease (Gupta *et al.,*

2004). Hence epitopes based vaccine can be developed which can be targeted against conserved regions of the viral proteins. Using experimental methods for epitope identification can prove to be expensive and time consuming and this is where an immunoinformatics driven approach for identification of epitopes against T and B cell epitopes comes into picture. Potential epitopes can be identified by major histocompatibility complex (HLA) binding algorithms. The ability by which these epitopes bind with the HLA class I or II determines the extent of $CD4^+$ or $CD8^+$ T-cell response. The focus of the present study is the identification of conserved immunogenic peptides containing multiple T and B cell epitopes present in the glycoprotein of the Ebola virus that can cover all the different globally distributed strains. Further molecular docking studies were performed with identified immunogenic peptides to calculate their binding affinity with different class I and II HLA molecules. These peptide fragments containing overlapping epitopes can be targeted as candidate for generating a universal vaccine against Ebola.

Chapter 2 REVIEW OF LITERATURE

2.1 Ebola Virus: History of the disease

On 23rd March 2014, WHO declared the outbreak of new disease Ebola virus disease (EVD) which started in Guinea, (December 2013) but quickly spread to other West African countries like Sierra Leone, Liberia and Nigeria. The first cases of Ebola virus disease were reported 1976, where two outbreaks of this disease occurred simultaneously, one in Nzara, Sudan infected by *Sudan Ebola virus* and other in Yambuku, Democratic republic of Congo infected by *Zaire Ebola Virus*. The third form of Ebola virus was found in 1989, when monkeys imported from Philippines into Virginia were infected by *Reston Ebola virus* (Saeidnia *et al.,* 2014*).* After Ebola Virus disease occurred in Africa in 1976-1979, it was not seen again until 1994.It is suspected that during those 15 years the virus was circulating in its own natural reservoir. In 1994, *Tai forest Ebola virus* emerged after infecting a Chimpanzee. The recent 2014 Ebola outbreak is the largest in history and also the first epidemic of Ebola world has ever seen affecting primarily Guinea, Liberia and Sierra Leone. The fatality rate in the Ebola virus disease is very high (approaching 90%) because of which, Centers for disease control and prevention (CDC) has classified it as the potential bioterrorism agent of Category A.

As per the report of CDC, total confirmed cases of Ebola are 15,113 out of which total deaths are 11,232 as of 26th June 2015. The first travel associated case of Ebola was reported in United States on September 30th 2014.

A small but unrelated outbreak also occurred in Democratic Republic of Congo in August 2014.

2.1.1 Taxonomy of Ebola Virus:

Ebola Virus disease previously known as Ebola hemorrhagic fever is a rare and deadly disease occurring in humans caused by infection by Ebola virus belonging to the *Filoviridae* family.

- **Order:** *Mononegavirales*
- **Family:** *Filoviridae*
- **Genus**: *Ebolavirus*
- **Species** 1) *Zaire ebolavirus*
  2) Sudan ebolavirus
  3) Tai Forest ebolavirus
  4) Bundibugyo ebolavirus
  5) Reston ebolavirus

Filovirus are negative sense RNA virus which are non-segmented, enveloped and pleomorphic. There are five strains of Ebola Virus out of which four can infect Humans:

- Zaire ebolavirus (EBOV): was recognized in 1976, has highest fatality rates and is responsible for the largest outbreak of Ebola in West Africa in 2014.
- Sudan ebolavirus (SUDV): was also seen in 1970 and has around 50% fatality rate.
- Tai Forest (TAFV): was formerly known as Ebola Ivory Coast. It has been found in only one person, an ethologist working with deceased chimpanzees.
- Bundibugyo ebolavirus (BDBV): latest species of Ebola virus which emerged in 2007. It has fatality rate of around 30%.
- Reston ebolavirus (RESTV): is maintained in an animal reservoir in the Philippines and is not found in Africa. It caused an outbreak of lethal infection in macaques imported into the United States in 1989. There is evidence that *Reston ebolavirus* can cause asymptomatic infection in humans.

Ebola is zoonotic but its natural reservoir still remains unclear. It is suspected that the fruit bats of the Pteropodidae family are natural virus hosts. First human case in an outbreak of Ebola is acquired through close contact with the infected blood, secretions or other bodily fluid of the infected animal. Human to human infection can spread through infected or dead person's blood or body fluids (urine, saliva, sweat, feces, vomit, breast milk, and semen) through broken skin or mucous

membranes or with indirect contact with contaminated objects like needles and syringes; surfaces and materials like bedding and clothing. Outbreaks have also been fuelled by traditional burial practices, in which mourners have direct contact with the bodies of the deceased. The onset of symptoms during EVD is sudden and are difficult to diagnose as they are non specific and similar to other diseases.

2.1.2 Structure of Ebola Virus:

Ebola Virus genome is 19kb long and consists of consists of seven genes: Nucleoprotein (NP), which is the most abundant protein in the viral cell, Glycoprotein (GP) which is the only surface protein, VP35 VP40, VP30, VP24 and polymerase genes Transcription is initiated by the RNA dependent RNA polymerase L, Polymerase Co-factor VP35 and transcription co-factor VP30 as shown in figure 1 (Brauburger *et al.*, 2014).

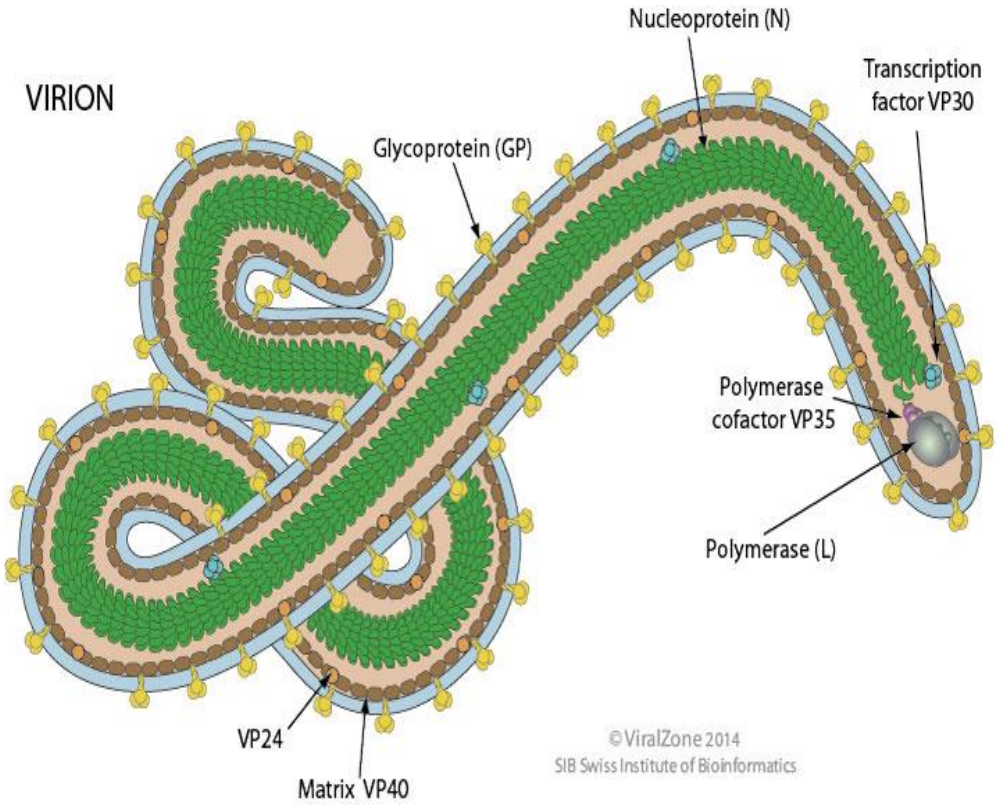


Figure 1: Structure of Ebola virus

Glycoprotein: Responsible for penetration of the virus into the cell cytoplasm by mediating the fusion of the membrane of the endocytosed virus particle with the endosomal membrane.

Nucleoprotein: Encapsidates the genome, protecting it from nucleases. The encapsidated genomic RNA is termed the nucleocapsid and serves as template for transcription and replication. During replication, encapsidation by NP is coupled to RNA synthesis and all replicative products are resistant to nucleases.

VP35: Acts as a polymerase cofactor in the RNA polymerase transcription and replication complex.

VP40: Promotes virus assembly and budding

VP30: Acts as a transcription anti-termination factor immediately after transcription initiation

VP24 : Prevents the establishment of cellular antiviral state by blocking the interferon-alpha/beta (IFN-alpha/beta) and IFN-gamma signaling pathways

RNA polymerase: is a RNA dependent RNA polymerase which carries out the replication of the genetic material.

2.1.3 Life Cycle of Ebola Virus:

Life cycle of Ebola virus can be divided into five stages as shown in figure 2:

1. Entry of virus particle into the host cell.

2. Entry of viral Ribonucleoprotein (RNP) into the nucleus.

3. Transcription and replication of the viral genome.

4. Export of the vRNPs from the nucleus.

5. Assembly and budding at the host cell plasma membrane

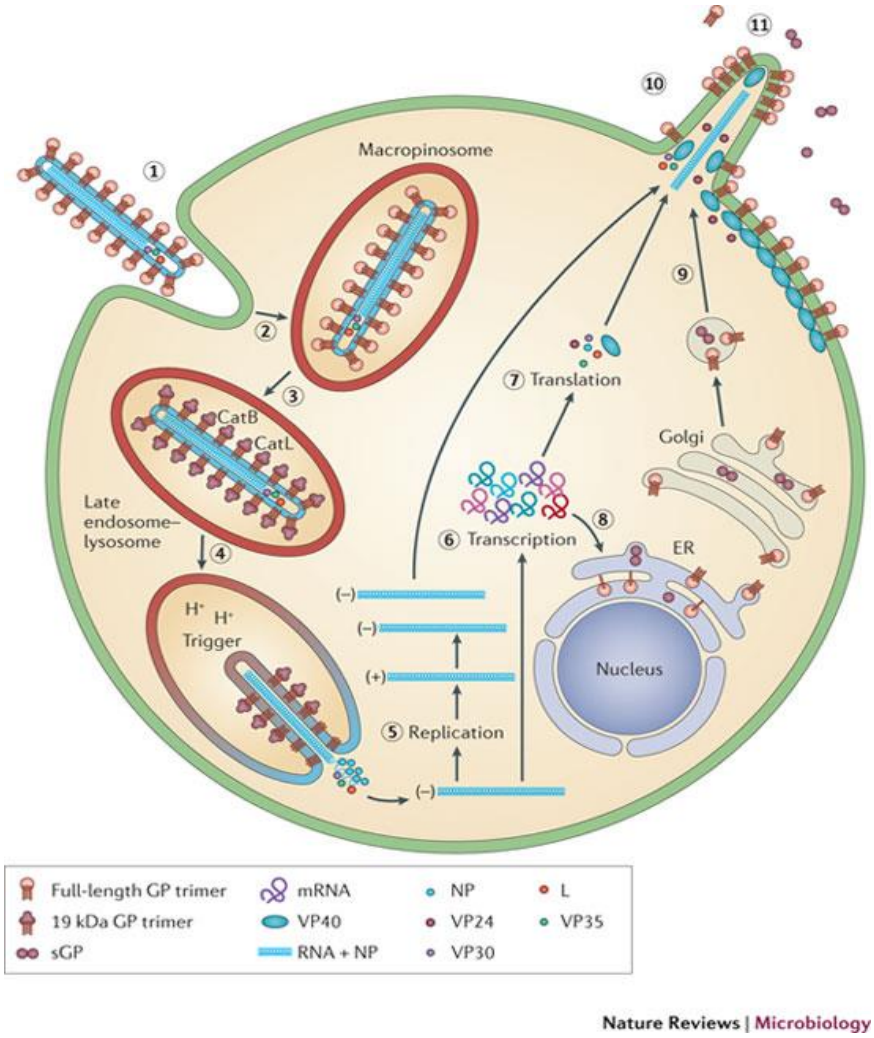


Figure 2: Life cycle of Ebola virus

Attachment to the host receptors is mediated by the spike protein; glycoprotein (GP).The virion enters the cytoplasm of the host cell through macropinocytosis. In the endosomal compartment GP protein is cleaved by cysteine proteases Cathepsin B and Cathepsin L into 19kb GP1.

GP1 mediated fusion of virus membrane with host cell and ribonucleoprotein is released into the cytoplasm which is followed by sequential transcription and replication. Nucleoprotein, viral polymerase, transcription activator VP20 and polymerase co-factor VP35 are together referred to as ribonucleoprotein complex [RNP Proteins] (Watt *et al.*, 2014).

2.1.4 Glycoprotein:

EBOV GP protein is the only virion surface protein and hence can prove a critical component for vaccines, inhibitors and neutralizing antibodies.

GP gene after transcriptional editing expresses two gene products: a 676-residues transmembrane linked glycoprotein (GP) and a 364-residues secreted glycoprotein (sGP).

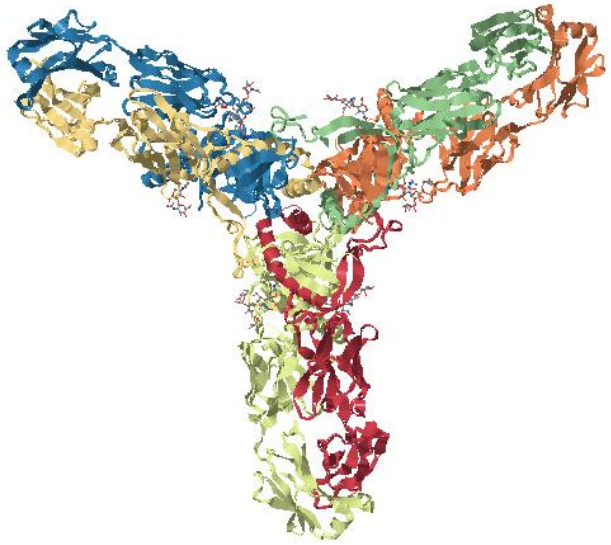


Figure 3: **Structure of crystal structure of the trimeric perfusion Ebola Virus glycoprotein in complex with a neutralizing antibody from a human survivor (PDB ID 3CSY)**

EBOV sGP is a non structural secreted protein which is suspected to play an important role in invasion of host immune response by inhibiting the activation of Neutrophils. It is also suggested that it has an anti-inflammatory response. sGP is also know to cross react with the antibodies secreted against the GP and neutralize them.

EBOV GP on the other hand is responsible for the attachment, fusion and entry of viral cell into the host cell membrane. While sGP is a dimer, GP is a trimer. GP

is cleaved post-translationally by Furin into GP_1 and GP_2 which are disulphide linked. Trimer of $GP_{1,2}$ form spikes present on viral surface. EBOV GP is a trimer which consists of three monomers which are non-covalently attached and form a Chalice like shape. Each monomer is found to consist of GP1 and GP2 (Lee *et al.*, 2008) Studies of the crystal structure also reveals that most of GP is shielded by a thick cloak of carbohydrate and identifies the very few sites left exposed and available for antibody binding, making this structure suitable as a template for vaccines and antibodies

EBOV GP1 is responsible for cell surface attachment to host cells which is probably mediated by residues 54 to 201. It is divided into head, tail and glycan cap region.

EBOV GP2 mediates fusion with host membranes and consists of two heptad repeat regions HR1 and HR2 and an internal fusion loop (Lee *et al.*, 2008)

During EBOV infection, virus encoded glycoprotein are released in the host cells in soluble form. High levels of truncated surface GP and sGP can be detected in the blood of infected patients. Primary infection starts with the replication of the virus in the dendritic cells, macrophages and monocytes followed by replication in hepatocytes and splenocytes. This is followed by massive release of cytokines, chemokines and vasoactive substances leading to inflammation disorders (Escudero-Pérez B et al., 2014). Glycoprotein binds to the endothelial cells leading to cytokine dysregulation and loss of vascular integrity leading to endothelial leakage and vasomotor collapse leading to rapid death of the host. Figure 4 shows the host cells response to the Ebola virus infection.

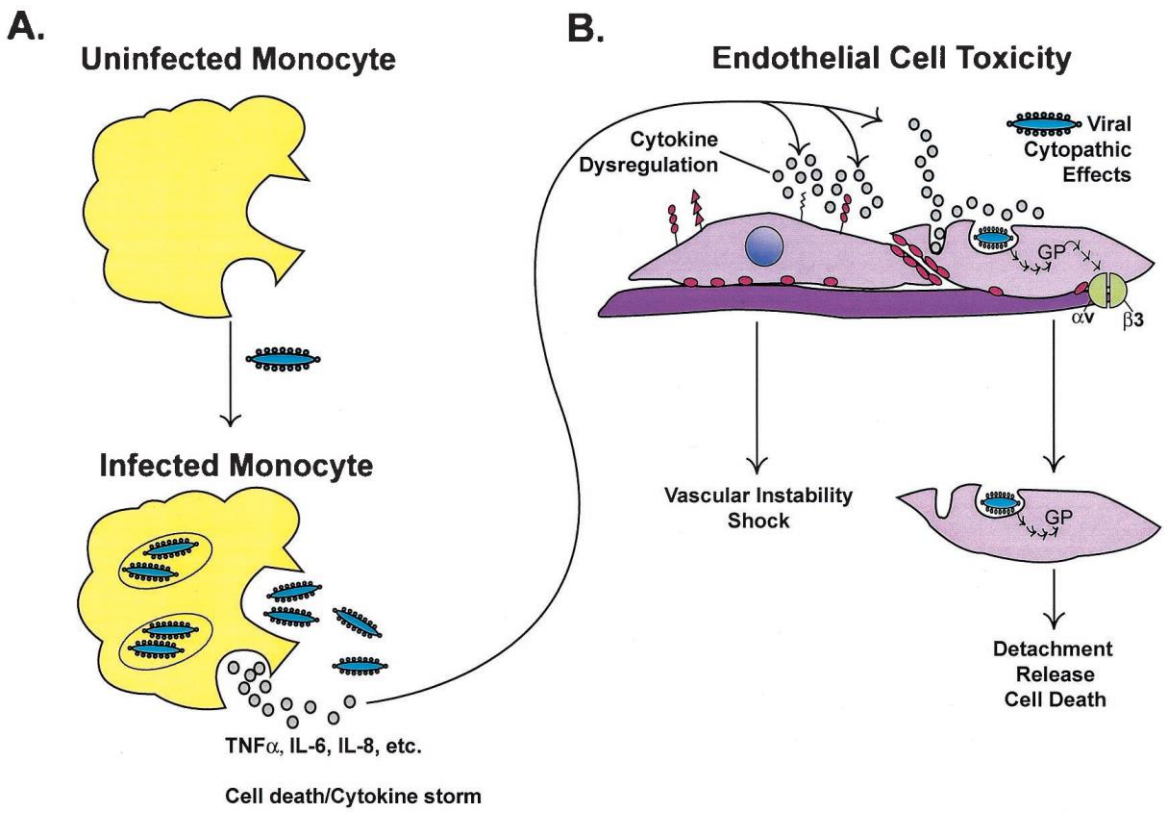


Figure 4: Infection of endothelial cells results in damage to endothelial barrier and leads to loss of vascular integrity. Cytokine dysregulation may synergize with the virus infection to lead hemorrhage. (Sullivan et al., 2003)

Evidence suggests that the Ebola virus Glycoprotein (GP) plays an important role in generating cytotoxic effects leading to cell damage by inducing cell rounding, detachment and membrane permeabilization. While sGP is suspected to alter immune response by inhibiting the activation of Neutrophils, GP contribute in the manifestations of Hemorrhage symptoms. (Sullivan, *et al*., 2013)

2.1.5 Mutation in Ebola Virus

Since discovered, Ebola has proven to be a stable virus with a relatively constant mutation rate. The Ebola virus samples from 2014 outbreak are 97% similar to the virus that first emerged in 1976. Scientists monitoring the virus have not seen any

evidence to suggest that the Ebola virus may be mutating to become more contagious or more easily spread.

2.1.6 Treatment for Ebola

There is no yet proven treatment available for Ebola and the rapid diagnostic kits for the early diagnosis of the disease are also lacking. Till now the treatment is only supportive However, a variety of vaccines, antibodies, small-molecule agents, and antiviral agents are undergoing testing.

ZMapp is one of the experimental drug which has shown some effective results against the disease. This drug is a cocktail of monoclonal antibodies with selected composition of c1326 from MB-003(human-mouse chimera mAbs and from ZMab (2G4 and 4G7).Both MB-003 and ZMab are also cocktails of antibodies. The components of ZMapp are produced from a tobacco plant (*Nicotiana benthamiana).* ZMapp was designed and produced through two major steps: first of all, the mouse was inoculated with Ebola virus, and then the immunological memory on the mouse lymphocytes by the infection was cloned. As a second part, cloned genes containing the memory related to Ebola virus was transferred to the tobacco plants by transformation using bacterial carrier. Finally plants expressing anti-Ebola mAbs were cultured in the fields and the antibodies were purified after harvest. ZMapp was supplied and helped to rescue the first American Ebola-infected patients this summer; however, another patient in Spain was not recovered even the ZMapp was treated. This drug began its human trial in Liberia in January 2015.

2.2 Vaccine Development for Ebola

Because the Ebola virus did not frequently generate outbreak, vaccine research and development was not eagerly preceded. Even though Ebola vaccine studies continued for last decade, the progress was not so fast or productive because of many limitations about the safety facilities and personal expertise. It is very difficult for the private research groups or companies to handle the selected agents or maintain the high safety level facilities, and usually the most steps of research were performed by limited agencies. Recently, some vaccine approaches using

DNA vaccine type containing Ebola antigen-coding genes announced positive results in animals.

cAd3 Ebola vaccine is a recombinant DNA vaccine which uses the adenovirus vector from a Chimpanzee. This vaccine aims at the expression of the GP gene in the host cell after vaccination. There are two vaccine candidates: Monovalent vaccine against ZEBOV and the Bivalent Vaccine against ZEBOV and SEBOV. This vaccine is undergoing human clinical trials.

Recombinant vesicular stomatitis virus Ebola vaccine is another DNA vaccine which is against Ebola as well as the Marburg Virus of the same family. This vaccine consists of highly attenuated recombinant form of VSV with GP of Ebola virus. This vaccine too is undergoing clinical trials in 2015.

2.2.1 Peptide Vaccine

The general concept behind the peptide vaccines is based on the chemical approach to synthesize the identified B-cell and T-cell epitopes that are immunodominant and can induce specific immune responses. B-cell epitope of a target molecule can be conjugated with a T-cell epitope to make it immunogenic. In epitope based vaccine immune response induced by an immunogen is not directed against the whole molecule, but it is against a specific region (sequence) of protein which is an epitope. The first epitope-based vaccine was created in 1985. They introduced recombinant DNA and expressed epitopes against cholera in Escherichia coli. Thus, epitope-based vaccines can be constructed for T and B lymphocytes. Because these vaccines pose no risk of mutation or reversion as well very little risk of contamination by toxic or pathogenic substances. They are comparatively easy to produce than a conventional vaccine and are chemically stable.

However epitope-based vaccines can be long-range and broad-spectrum vaccines, but with the use of small molecules such as peptides, a problem of low immunogenicity is encountered, as compared to multi-epitope protein antigens, or the entire pathogen, that is used for immunization in conventional vaccines.

2.3 Immunoinformatics:

Immunoinformatics can be described as a branch of bioinformatics concerned with in silico analysis and modeling of immunological data and problems. Immunoinformatics research stresses mostly on the design and study of algorithms for mapping potential B- and T-cell epitopes, which speeds up the time and lowers the cost needed for laboratory analysis of pathogen gene products. Using such tools and information, an immunologist can analyze the sequence areas with potential binding sites, which in turn leads to the development of new vaccines. The methodology of analysing the pathogen proteome to identify the potential immunogenic protein is known as reverse vaccinology.

Conventional methods need to dedicate time to pathogen cultivation and subsequent protein extraction. Although pathogens grow quickly, extraction of their proteins and then testing of those proteins on a large scale is expensive and time-consuming. Immunoinformatics is capable of reducing time and saving resources for the development of relevant vaccines by directing towards the development of more rational peptide based vaccine approach.

2.4 MHC Polymorphism and Vaccine Design

The major histocompatibility complex is a collection of genes arrayed within a long continuous stretch of DNA on chromosome 6 in humans. The MHC is referred to as the HLA complex in humans. HLA genes are organized into regions encoding three classes of molecules. Class I HLA genes encode glycoprotein expressed on the surface of nearly all nucleated cells; the major function of the class I HLA is presentation of intracellular antigens to T cytotoxic cells. Class II HLA genes encode glycoprotein expressed primarily on antigen-presenting cells (macrophages, dendritic cells, and B cells), where they present processed extracellular antigenic peptides to T-cells. Class III HLA genes encode, in addition to other products, various secreted proteins that have immune functions, including components of the complement system and molecules involved in inflammation (Kuby Immunology, Chapter 6).

The loci constituting the HLA are highly polymorphic; that is, many alternative forms of the gene, or alleles, exist at each locus among the population. The genes of the HLA loci lie close together. HLA class I gene complex contains three different HLA loci A, B and C each of which codes of α chain polypeptides of HLA class I molecule. The class II gene complex contains three major HLA loci, DP, DQ and DR; each of these loci codes α and β chain polypeptides of HLA class II molecule. Most individuals inherit the alleles encoded by these closely linked loci as two sets, one from each parent. Each set of alleles is referred to as a haplotype. The alleles are codominantly expressed; that is, both maternal and paternal gene products are expressed in the same cells.

The HLA polymorphism within a species generates a diversity of binding specificities, and thus different patterns of responsiveness to antigens. So T cell epitopes which bind to multiple HLA molecules will be an effective vaccine target as it can provide better coverage of the population

2.4 T-cell and B-cell epitopes and their prediction algorithms:

Understanding the immune epitope recognition is critical for development of vaccines. Epitopes are defined as the part of the molecule that are specifically recognized by the immune system. They are divided into two types: B-cell receptors which are recognized by the B cell receptors (BCR) and the antibodies, and the T-cell epitopes which are presented by the Major Histocompatibility complex (HLA) and T-cell receptors (TCR).

T-cell epitope identification and selection is the most crucial step in epitope based vaccine development. T cells can recognize antigen only when it is presented by a group of specialized proteins known as HLA or HLA, class I and class II. Experimental methods of epitope prediction are expensive and time consuming hence *in silico* approaches are used for epitope prediction.

Cytotoxic T lymphocytes (CTLs) play an important role in the control of virus infections. Targets for CTL responses, such as the virus surface protein (GP) and the nucleoprotein (NP), are considered as candidate vaccines because of their

conserved nature. The use of conserved proteins could provide protective immunity against drift variants or viruses with novel subtypes.

HLA class II proteins bind oligopeptide fragments derived through the proteolysis of pathogen antigens, and present them at the cell surface for recognition by $CD4^+$ T cells. If sufficient quantities of the epitope are presented, the T cell may trigger an adaptive immune response specific for the pathogen. Class II HLAs are expressed on specialized cell types, including professional APCs such as B cells, macrophages and Dendritic cells, whereas class I HLAs are found on every nucleated cell of the body. The HLA I molecule binds to a peptide of approximate 9 amino acids in length within a closed groove. In contrast, because the antigen-binding groove is open at both ends, the HLA II molecules can present much longer peptides, generally varying from 12 to 25 amino acids, nine of which occupy the binding groove. This difference between HLA I and HLA II is very important for the development of distinct prediction algorithms (Larson *et al*., 2006).

One of the key issues in T-cell epitope prediction is the prediction of HLA binding, as it is considered a prerequisite for T cell recognition. All T-cell epitopes are good HLA binders, but not all good HLA binders are T-cell epitopes. Determining the peptide binding preferences exhibited by this extensive set of alleles is beyond the present capacity of experimental techniques, necessitating the development of bioinformatics prediction methodologies. The most successful prediction methods for T-cell epitopes developed to date have been data-driven. T-cell epitope prediction typically involves defining the peptide binding specificity of specific class I or class II HLA alleles and then predicting epitopes *in silico*. Using peptide sequence data, experimentally determined affinity data have been used in the construction of many T-cell epitope prediction algorithms. Such methods include motif-based systems, support vector machines (SVMs), and hidden Markov models (HMMs), quantitative structure–activity relationship (QSAR) analysis and structure-based approaches.

Compared to T cell epitopes prediction algorithms, the B cell epitope prediction is more complicated, especially for the conformational B cell epitopes because, in addition to the sequence composition, the 3D-structure of protein must also be considered.

The prediction algorithms for linear B are similar to that of T-cell. The accuracy of primary sequence based algorithm is low and modified algorithms based on machine learning were subsequently developed, such as ABCpred (Saha *et al.,* 2006) and BepiPred (Larsen *et al.,* 2006) with significant improvements in accuracy. Prediction algorithms for conformational B cell epitopes based on 3D structure are also available owing to the ever-increasing 3D structure of antigen-antibody complex data. Some prediction servers based on this algorithm are accessible, for example DiscoTope and CEP (http:// bioinfo.ernet.in/cep.htm) (Kulkarni *et al*., 2005). These methods make use of information carried in the structure of antibodies against proteins of interest to reveal the 3D folding of target proteins.

2.5 Peptide Vaccine and its potential as a Vaccine against Ebola:

Extremely lethal, Ebola virus requires Biosafety level 4 Containment (BSL-4), which hampers the study of live virus. A high degree of biohazard containment is required for laboratory studies and clinical analysis. These factors limit the usage of conventional research methods for vaccine development.

It has also been reported that $CD8^+$ cells play a significant role in initial clearance of Virus during acute infection. In the absence of $CD8^+$ cells, the immune system fails in controlling the virus replication. Hence $CD8^+$ cells play an important role in controlling the initial progression of the Ebola virus disease and also protects against fatality. B-cells and $CD4^+$ cells on the other hand, are required for long term protection and for avoiding the reemergence of the disease. (Gupta *et al.,* 2004).Hence epitopes based vaccine can be developed which can be targeted against conserved regions of the viral proteins.

Chapter 3: OBJECTIVES

- Identification of conserved peptides containing T and B cell epitopes of Ebola Virus Glycoprotein using different immunoinformatics tools.

- Population coverage and HLA distribution analysis of the predicted peptides to identify their capacity of inducing immune response among the different populations of the world

- Molecular docking of predicted peptides with the HLA class I and II molecules.

Chapter 4 MATERIALS AND METHODS

4.1 Sequence

The Ebola virus Glycoprotein consists of 676 amino acids. Glycoprotein sequences of ***Zaire ebolavirus, Sudan ebolavirus and Bundibugyo ebolavirus*** available till August 2014 were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/protein).

Complete protein sequences with Human as host were downloaded in FASTA format and saved in Microsoft office word. Partial or incomplete protein sequences were avoided.

4.1.2 Multiple sequence alignment

Multiple sequence alignment is carried out to align multiple related sequences to achieve optimal matching of the sequences. Multiple sequence alignment was carried out by MUSCLE tool. (http://www.ebi.ac.uk/Tools/msa/muscle)

MUSCLE stands for **M**ultiple **S**equence **C**omparison by **L**og- **E**xpectation. MUSCLE is a computer program for creating multiple alignments of protein sequences. Elements of the algorithm include fast distance estimation using kmer counting, progressive alignment using the log-expectation score, and refinement using tree dependent restricted partitioning. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee.

The limitation of MUSCLE is that it can only align 500 sequences of more than 350 amino acid length at a time. The number of sequence retrieved is more than 500 and MUSCLE (Multiple sequence alignment tool) can only align only 500 sequences at a time. The results of MUSCLE were saved in FASTA format.

4.1.3 Conservancy Analysis

Conservancy analysis was carried out to find out the conserved region present in glycoprotein of Ebola virus. The regions of glycoprotein showing ≥ 90% conservancy were selected. AVANA tool was used for conservancy analysis.

AVANA

The Antigen Variability Analyzer (AVANA) tool uses information entropy to measure variability in protein sequence alignments. It also compares alignments using mutual information, identifying the mutations that characterize specific sequence sets (http://avana.sourceforge.net/).

Alignment result of MUSCLE (FASTA format) was used as input for AVANA software. Parameters were set to 90% conservancy and minimum length of 9 amino acids as a threshold value in AVANA. Conserved regions were searched in the alignments. The overlapping regions found in the conserved regions were merged together.

4.2 $CD8^+$ T-cell Epitope prediction

$CD8^+$ cells recognise the peptides presented on class I HLA by the cells which are either tumorous in nature or viral infected. The target cells are killed by the release of perforins and granzymes by $CD8^+$ cells. For the identification of $CD8^+$ specific epitopes three different tools were used.

4.2.1 BIMAS (http://www-bimas.cit.nih.gov/molbio/hla_bind/)

BIMAS stands for Bioinformatics and Molecular Analysis Section. BIMAS locate and rank 8 mer, 9-mer, or 10-mer peptides that contain peptide-binding motifs for HLA class I molecules. It predicts the potential epitopes for 33 HLA Class I alleles on the basis of half time of the dissociation from the β2 microglobulin of Class I HLA molecules. The tool predicts the binding of the peptide to the HLA from the available peptide binding data of 152 peptides which depends upon their respective position on the epitope.

Eighty peptides with binding half-life of binding with β2 microglobulin more than 5 min has been included in the algorithm for the prediction of potential epitopes similarly the remaining 72 peptides which had half-life less than 5 min used to eliminate the false prediction (Parker *et al.,* 1994). The cut-off of 50 was used for BIMAS for the epitope prediction. It may be helpful for the prediction of stable and potential epitopes.

4.2.2 SYFPEITHI

SYFPEITHI makes the predictions for both class I and class II HLA alleles of human, mouse, rat, ape, cattle, and chicken. The algorithm identifies the anchor and auxiliary residues in the input sequence by comparing it with the database of naturally occurring epitopes. It assigns particular score to each residue depending upon their position and generate overall score for predicted epitope (Rammensee *et al*., 1999). The score is calculated by assigning amino acids of a certain peptide a specific value depending on the fact if they are anchor, auxiliary anchor or preferred residue. Anchors are given 10 points; unusual anchors 6-8 points, auxiliary anchors 4-6 and preferred residues 1-4 points. The amino acids which are responsible for hindering the binding ability are assigned score between -1 and -3. The cut off value taken was 20 for the predicted epitopes.

4.2.3 NetCTL 1.2

NetCTL 1.2 leads to the epitope prediction on the basis of C-terminal proteasomal cleavage, TAP transport efficiency and HLA binding. It identifies the epitopes for 12 HLA-A and HLA-B super types (Larsen *et al*., 2007). The cut off value >0.75 was used for the epitope prediction. The cut off 0.75 signifies the 80% sensitivity and 0.97 specificity score of the predicted epitopes. The three combined scores result in the identification of the potential immunogenic epitopes from the given sequence. The common epitopes predicted by all the three tools were selected. The selected epitopes were overlapped to generate the peptide fragments which consisted of multiple epitopes.

4.3 $CD4^+$ T-cell Epitope prediction

Class II HLA molecules bind peptides and present these peptides to $CD4^+$ T cells. Peptides of class II HLA–peptide complexes is generally 13–18 amino acid residues long, somewhat longer than the nonameric peptide of class I molecules but the core sequence is still 9 amino acid in length. The peptide-binding cleft in class II molecules is open at both ends, allowing longer peptides to extend beyond the ends. Three different online software having different epitope prediction algorithms were used to identify HLA class II epitope To identify the $CD4^+$

specific epitope three different tools Rankpep (Reche *et al*., 2007), Propred (Singh *et al*., 2001) and IEDB consensus (Wang *et al.,* 2008) were used.

4.3.1 Rankpep

It predicts the Class II binding peptides from already known class II HLA binding aligned peptides by using PSSM (Position Specific Scoring Matrix). (Reche *et al.*, 2007). The HLA binding score of the peptide is obtained by aligning the PSSM with the protein segment and assigning the score to the residue which matches to the profiles in the matrix. The threshold value for the epitope prediction was taken top 2% of the predicted peptides.

4.3.2 Propred (http://www.imtech.res.in/raghava/propred/)

ProPred is a graphical online tool for predicting HLA class II binding regions in antigenic protein sequences. The server implement Quantitative Matrix based prediction algorithm, employing amino-acid / position coefficient table deduced from literature. The predicted binders can be visualized either as peaks in graphical interface or as colored residues in HTML interface (Singh H. *et al., 2001*). The threshold value for the epitope prediction was taken top 2% of the predicted peptides.

ProPred measures the percentage score with respect to the best score for that particular allele. In other words if score of a epitope is less than 2% of the best score for that specific HLA allele then it will be considered as a negative result.

4.3.3 IEDB consensus

The IEDB consensus apply the multiple algorithms, NN-align, SMM-Align and combinatorial method for the prediction of available epitopes otherwise uses the netHLApanII. The performance for HLA-II binding methods are in order: Consensus > NetHLAIIpan > NN-align > SMM-align > Combinatorial Library. Based upon the epitope predicted by different algorithms a percentile rank is generated for the finally predicted epitopes (Wang *et al*., 2008). The threshold

values taken for IEDB consensus were percentile rank of top 10% of predicted epitopes. The common epitopes predicted by all the three tools were selected.

4.4 Blast Screening

In order to avoid any similarity of the peptides with functional human protein other than Ebola Glycoprotein, BLAST analysis (Altschul *et al*., 1990) was performed for epitopes predicted to bind HLA class I and II respectively. The peptides showing similarity in amino acids without gap or mismatch were eliminated (Tan *et al*., 2010) thus ruling out any possibility of autoimmune response against any human functional protein. The peptides screened after BLAST showing overlapped and merged together to generate single peptide fragment.

4.5 Population Coverage Analysis

HLA alleles are highly polymorphic and diversity of alleles are found in the global population. Different individuals of the worldwide population may respond in different way against particular antigen. The IEDB database enables predicting the possible world population capable of responding to particular immunogen. The database contains variety of HLA class I and II alleles frequently found in different world population of geographical areas. It compares the epitopes specific alleles with the alleles in database and find out percentage population capable of responding. In our study IEDB population coverage analysis tool was carried out for 16 different geographical areas distributed globally (Huynh-Hoa Bui *et al.*, 2006).

4.6 Epitope Docking

4.6.1 Peptides Structure Prediction

The structure of predicted epitopes common for class I and Class II HLA were predicted using peptide structure prediction server PEP-FOLD (Thevenet *et al*., 2012). The model generated by PEPFOLD are assorted either by using the coarse grain energy of the PEPFOLD or by predicted Tm score. For the peptides of length 36 residues the coarse grain energy of the PEPFOLD is used to sort the

predicted models whereas Tm score for the Peptides with the residues more than 36.

4.6.2 Separation of naturally bound peptide from HLA molecule.

The structure of one peptide bound HLA class I and one HLA class II were retrieved. The ligands and receptor of the PDB structure were separated using the discovery studio visualization tool 3.5. The generated peptide free HLA structures were further used for docking using Auto dock Vina (Trott *et al*., 2009) tool.

Ligand Separation

Open the PDB structure of HLA molecule in discovery studio visualization tool 3.5

⬇

Go to the scripts option and select the water molecules

⬇

Go to edit and delete

⬇

Go to scripts and select protein chains

⬇

Go to edit and delete

⬇

Save as ligand.pdb file

Receptor Separation

Open the PDB structure of HLA molecule in discovery studio visualization tool 3.5

⬇

Go to the scripts option and select the water molecules

⬇

Go to edit and delete

⬇

Go to scripts and select ligands

⬇

Go to edit and delete

⬇

Save as the receptor.pdb file

Receptor preparation for docking

Open the PDB structure of receptor molecule in Auto dock Vina 4.0

⬇

Go to the edit and add hydrogen atoms (polar only)

⬇

Go to grid and choose and select the macromolecule i.e. receptor molecule

⬇

Go to grid box and select the grid where the Ligand should be bound

⬇

Save the file as receptor.pdbqt

Ligand preparation for docking

Open the ligand.pdb file in the Auto dock Vina 4.0

Go to Ligand and choose the molecule

Select the Ligand molecule

Go to output and save the molecule as ligand.pdbqt

4.6.3 Preparation of configuration file

To perform the docking in auto dock vina we are supposed to generate a text file named as config containing the following commands. The config file should be created for the each molecule separately depending upon their grid box selection.

receptor = receptor.pdbqt
ligand = ligand.pdbqt
out = out.pdbqt

center_x = 40.152 (molecules specific obtained by grid box selection)
center_y = -10.429 (molecules specific obtained by grid box selection)
center_z = - 26.057 (molecules specific obtained by grid box selection)

size_x = 56 (molecules specific obtained by grid box selection)

size_y = 70(molecules specific obtained by grid box selection)

size_z = 48(molecules specific obtained by grid box selection)

Exhaustiveness = 16.

4.6.4 Docking

Molecular docking is a key tool in structural molecular biology and computer-assisted drug design. The goal of ligand-protein docking is to predict the predominant binding mode(s) of a ligand with a protein of known three-dimensional structure. Successful docking methods search high-dimensional spaces effectively and use a scoring function that correctly ranks candidate dockings.

The separated naturally bound peptides of the HLA molecules were redocked with the HLA peptide free HLA using AutoDock vina (Trott *et al*., 2009) tool. The binding energy of these naturally bound ligands of HLA molecules was determined. The predicted epitopes were docked with the HLA class I and II structure and binding energy was compared with the binding energy of the redocked naturally bound peptides. The docking was performed at exhaustiveness 16 with all the peptides and HLA structures.

4.7 Prediction of B-cell specific linear epitope

Antibodies generated in response to an immunogen on the cell surface leads to cell killing by mean of antibody dependent cell mediated cytotoxicity (ADCC). A humoral immune response against cancer antigen can help in cancer cell killing by mean of same mechanism. To identify the B-cell specific linear epitope IEDB Kolaskar and Tongaonkar (Kolaskar *et al*., 1990) method was used, using default threshold value 0.992 and window size 7.

This method apply the semi empirical approach to identify the antigenic determinant on protein depending upon their physicochemical properties and frequency of occurrence of amino acids on it based upon the previously known B-cell epitopes, The tool predicts the antigenic propensity of the predicted peptides.

The fundamental of the determining the antigenicity is that if cysteine, leucine and valine occur on the surface of the protein they are more likely to act as B-cell antigenic site.

The percentage accuracy of this method is found out be nearly 75% which is better than any other known B-cell epitope prediction method (Kolaskar *et al*., 1990). The predicted B-cell epitopes were further compared with the already predicted T-cell epitopes explained above, to find out if any similarity between T-cell and B-cell linear epitopes exists.

Chapter 5 RESULTS AND DISCUSSION

5.1 Conserved regions of GP in Ebola Virus

For the purpose of covering all the genetic variants of the three virulent strains: ***Zaire ebolavirus, Sudan ebolavirus and Bundibugyo ebolavirus*** isolates present all over the world, 159 Glycoprotein sequences were obtained. Sequences were aligned using the MUSCLE tool and conserved regions were predicted using the AVANA. Twelve peptide fragments (GSP1-GSP12), ranging from 13 to 136 amino acids were found to be conserved (Table 1). These 12 fragments were then selected for epitope prediction.

Table1: CONSERVED PEPTIDE FRAGMENTS OF EBOLA VIRUS GLYCOPROTEIN

S. No	Conserved peptide sequence	Length
1	TGILQLPRDRFKRTSFFLWVIILFQRTFSIPLG	33
2	VDKLVCRDKLSSTNQLR	17
3	SVGLNLEGNGVATDVPS	16
4	TKRWGFRSGVPPKVVNYEAGEWAENCYNLEIKKPDGSECLPAAPDGIRGFPRC RYVHKVSGTGPCAGDFAFHKEGAFFLYDRLASTVIYRGTTFAEGVVAFLILPQ AKKDFFSSHPLREPVNATEDPSSGYYSTTI	136
5	YQATGFGTNETEYLFEVDNLTYVQLE	26
6	RFTPQFLLQLNET	13
7	SGKRSNTTGKLIWKVNPEID	20
8	GNNNTHHQDTGEES	14
9	NTIAGVAGLITGGRR	15
10	IVNAQPKCNPNLHYWTTQDEGAA	23
11	KCNPNLHYWTTQDEGAAIGLAWIPYFGPAAEGIYTEGLMHNQ	34
12	GLICGLRQLANETTQALQLFLRATTELRTFSILNRKAIDFLLQRWGGTCHILGP DCCIEPHDWTKNITDKIDQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGV TGVIIAVIALFCICKFV	123

5.2 $CD8^+$ T-cell specific epitope identification

$CD8^+$ T cells are often called as cytotoxic T lymphocytes (CTL) and are important for immune defense against intracellular pathogens including virus and bacteria and for tumor surveillance. $CD8^+$ cells recognise antigens which are associated with HLA class I molecules, found in all nucleated cells.

Three tools used for the prediction of HLA class I restricted epitopes were NetCTL, BIMAS and SYFPEITHI. Peptides were considered as epitopes only when they were commonly predicted by all three tools. Out of 28 predicted epitopes, seven were found to be homologous to human proteins after BLAST screening and were eliminated. Final result consisted of 21 $CD8^+$ T-cell epitopes which were selected for future consideration. (Table 2)

The epitopes that had overlapping amino acid sequences were merged to form a single peptide fragment. Eleven peptide fragments containing overlapping $CD8^+$ T-cell epitopes were generated, ranging from nine to thirty four amino acids as listed in Table 3.

Table 2: $CD8^+$ T-cell specific Epitopes

Peptide	$CD8^+$ specific T-cell Epitopes	Amino Acid Position
GSP1	DRFKRTSFF	12
GSP1	RFKRTSFFL	13
GSP1	KRTSFFLWV	14
GSP1	FQRTFSIPL	25
GSP4	FLYDRLAST	160
GSP4	DRLASTVIY	163
GSP4	AEGVVAFLI	176
GSP4	ATEDPSSGY	214
GSP5	EYLFEVDNL	231
GSP5	YLFEVDNLT	232
GSP5	FEVDNLTYV	234
GSP7	KRSNTTGKL	265
GSP11	AEGIYTEGL	549
GSP12	LANETTQAL	572
GSP12	NETTQALQL	574
GSP12	LFLRATTEL	582
GSP12	RTFSILNRK	591
GSP12	NRKAIDFLL	597
GSP12	LRTFSILNR	590
GSP12	IHDFVDKTL	638
GSP12	ALFCICKFV	678

Table 3: Peptide containing overlapping $CD8^+$ specific T-cell epitopes

Peptide Enriched $CD8^+$ specific T-cell epitopes	**Number of Epitopes**
DRFKRTSFFLWV	3
FQRTFSIPL	1
FLYDRLASTVIY	2
AEGVVAFLI	1
ATEDPSSGY	1
EYLFEVDNLTYV	3
KRSNTTGKL	1
AEGIYTEGL	1
LANETTQALQLFLRATTELRTFSILNRKAIDFLL	6
IHDFVDKTL	1
ALFCICKFV	1

5.2 $CD4^+$ T-cell specific epitope identification

$CD4^+$ T-cells are often called as helper T-cells as after being activated they play a major role in immune defense by secretion of specific cytokines. $CD4^+$ T-cells recognise the antigens present on HLA class II molecules and once activated, divide rapidly to activate other cells of the immune system like $CD8^+$ T-cells and B cells.

For prediction of HLA class II epitopes three tools, Rankpep, Proped and IEDB recommended methods were used. Peptides were considered as epitopes only when they were commonly predicted by all three tools. Out of 13 predicted epitopes, two were found to be homologous to human proteins after BLAST

screening; these epitopes were eliminated. Finally eleven epitopes were taken for future consideration (Table 4).

The epitopes that had overlapping amino acid sequences were merged to form a single peptide fragment. Hence seven overlapping peptide fragments were obtained which contained one or more than one $CD4^+$ T-cell epitope (Table 5)

Table 4: $CD4^+$ specific T cell Epitopes

Peptide	**$CD4^+$ specific T-cell Epitopes**	**Amino Acid Position**
GSP1	ILFQRTFSI	25
GSP1	FKRTSFFLW	14
GSP2	LVCRDKLSS	48
GSP4	IRGFPRCRY	129
GSP4	YVHKVSGTG	137
GSP4	FHKEGAFFL	152
GSP4	FFLYDRLAS	159
GSP12	LRQLANETT	570
GSP12	FLRATTELR	584
GSP12	LFLRATTEL	583
GSP12	WTKNITDKI	627

Table 5: Peptide containing overlapping $CD4^+$ specific T-cell epitopes

Peptide Enriched $CD4^+$ specific T-cell epitopes	**Number of Epitopes**
FKRTSFFLW	1
ILFQRTFSI	1
LVCRDKLSS	1
IRGFPRCRYVHKVSGTG	2
FHKEGAFFLYDRLAS	2
LRQLANETT	1
LFLRATTELR	2
WTKNITDKIDQ	1

5.3. Identification of Peptide containing $CD4^+$ and $CD8^+$ specific epitope

It was interesting to find peptides that contained both $CD4^+$ as well as $CD8^+$ T-cell epitopes. Hence identification of those conserved peptide fragments were done which contained overlapping $CD4^+$ and $CD8^+$ T-cell epitopes. This resulted in generation of four conserved peptide fragments that contained both $CD4^+$ and $CD8^+$ T-cell epitopes (Table 6). These peptide fragments (peptide 1 to peptide 4) consist of twelve $CD8^+$ T-cell epitopes and six $CD4^+$ T-cell epitopes which are listed in table 7.

Table 6: Ebola Glycoprotein peptides containing overlapping $CD8^+$ and $CD4^+$ specific T-cell Epitopes

Peptide enriched $CD8^+$ specific T-cell epitopes	No. of Epitopes	Peptides enriched $CD4^+$ specific T-cell epitopes	No. of Epitopes	Peptide containing $CD8^+$ and CD4+ specific T-cell epitopes
DRFKRTSFFLWV	3	FKRTSFFLW	1	DRFKRTSFFLWV
FQRTFSIPL	1	ILFQRTFSI	1	ILFQRTFSIPL
FLYDRLASTVIY	2	FHKEGAFFLYDRLAS	2	FHKEGAFLYDRLASTVIY
LANETTQALQLFLRATTELRTFSILNRKAIDFLL	6	LRQLANETT	1	LRQLANETTQALQLFLRATTELRTFSILNRKAIDFLL
	1	LFLRATTELR	2	

Table 7: Final Ebola Glycoprotein peptides containing twelve $CD8^+$ T cell epitopes (G1 to G12) and six $CD4^+$ T-cell Epitopes

Conserved Peptide Fragment	CD8+ Tcell epitopes	CD4+Tcell epitopes
DRFKRTSFFLWV	DRFKRTSFF KRTSFFLWV RFKRTSFFL	FKRTSFFLW
ILFQRTFSIPL	FQRTFSIPL	ILFQRTFSI
FHKEGAFLYDRLASTVIY	FLYDRLAST DRLASTVIY	FFLYDRLAS
LRQLANETTQALQLFLRATTELRTFSILNRKAIDFLL	NRKAIDFLL ALFCICKFV LRTFSILNR LANETTQAL NETTQALQL LFLRATTEL	LRQLANETT FLRATTELR LFLRATTEL

5.4. Population coverage and HLA distribution Analysis

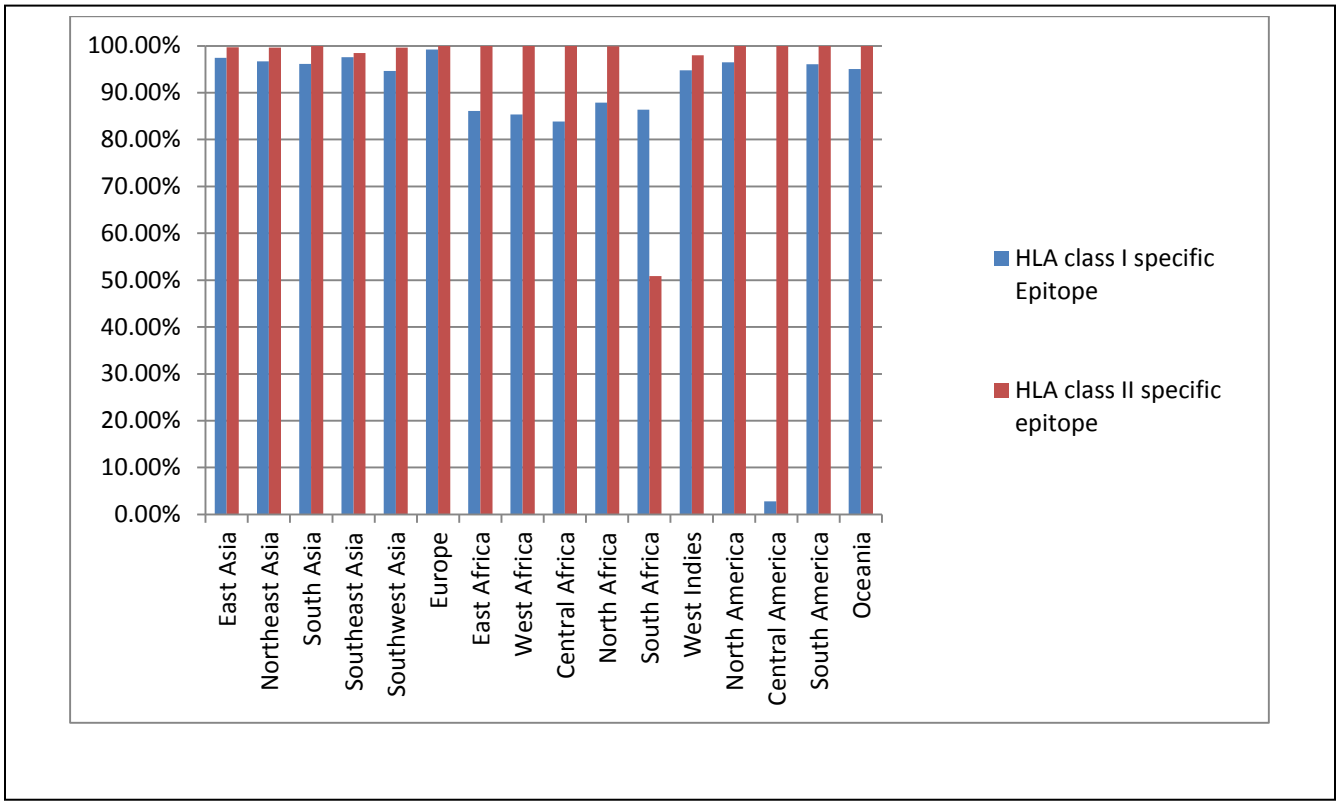
HLA distribution varies among different ethnic groups around the globe. Population coverage analysis plays an important role for the design of peptide-based vaccine. Due to highly polymorphic nature of HLA, the response against a particular antigen throughout a human population is restricted. Hence a Tcell epitope which binds to several HLA alleles is preferred for maximum population coverage. The peptide fragments were found to cover larger number of HLA class I and II alleles (Table 8).

In order to estimate the immunogenic potential of the predicted peptide fragments on the global population, population coverage analysis was carried out using the IEDB population coverage tool. IEDB tool calculated the expected fraction of individuals from different populations that were going to respond to the peptide fragments. The average population coverage for immunogenic response of predicted peptides were found out to be 87.28% and 96.63 % for class I and II HLA specific respectively (Figure 5).

Table 8: Number of HLA class I and II alleles capable of responding against peptide fragment containing both $CD4^+$ and $CD8^+$ specific T-cell epitope

Peptide Fragment	Number of T- cell specific HLA Class I specific Alleles	Number of T- cell specific HLA Class II specific Alleles
DRFKRTSFFLWV	30	66
ILFQRTFSIPL	30	66
FHKEGAFLYDRLASTVIY	25	55
LRQLANETTQALQLFLRATTELRTFSILNRKAIDFLL	37	74

Figure 5: Population coverage analysis of peptides containing $CD4^+$ and $CD8^+$ T cell epitopes for class I and II HLA alleles in 17 different geographic regions predefined in the IEDB database.



5.5 Peptides Structure Prediction

PEP-FOLD predicts the structure from the amino acid sequence. It is based upon structure alphabet letters to describe the structure of the given sequence. It predicts the structure of four consecutive amino acids and then combines the series of structural alphabet letter to generate the structure (Thevenet *et al.*, 2012). PEPFOLD was used for prediction of all four peptide fragments containing $CD4^+$ and $CD8^+$ specific T-cell epitopes as well as for 12 epitopes specific for $CD8^+$ T cells (Listed in Table 7). The best model generated was taken (Figure 6) and subjected for epitope docking.

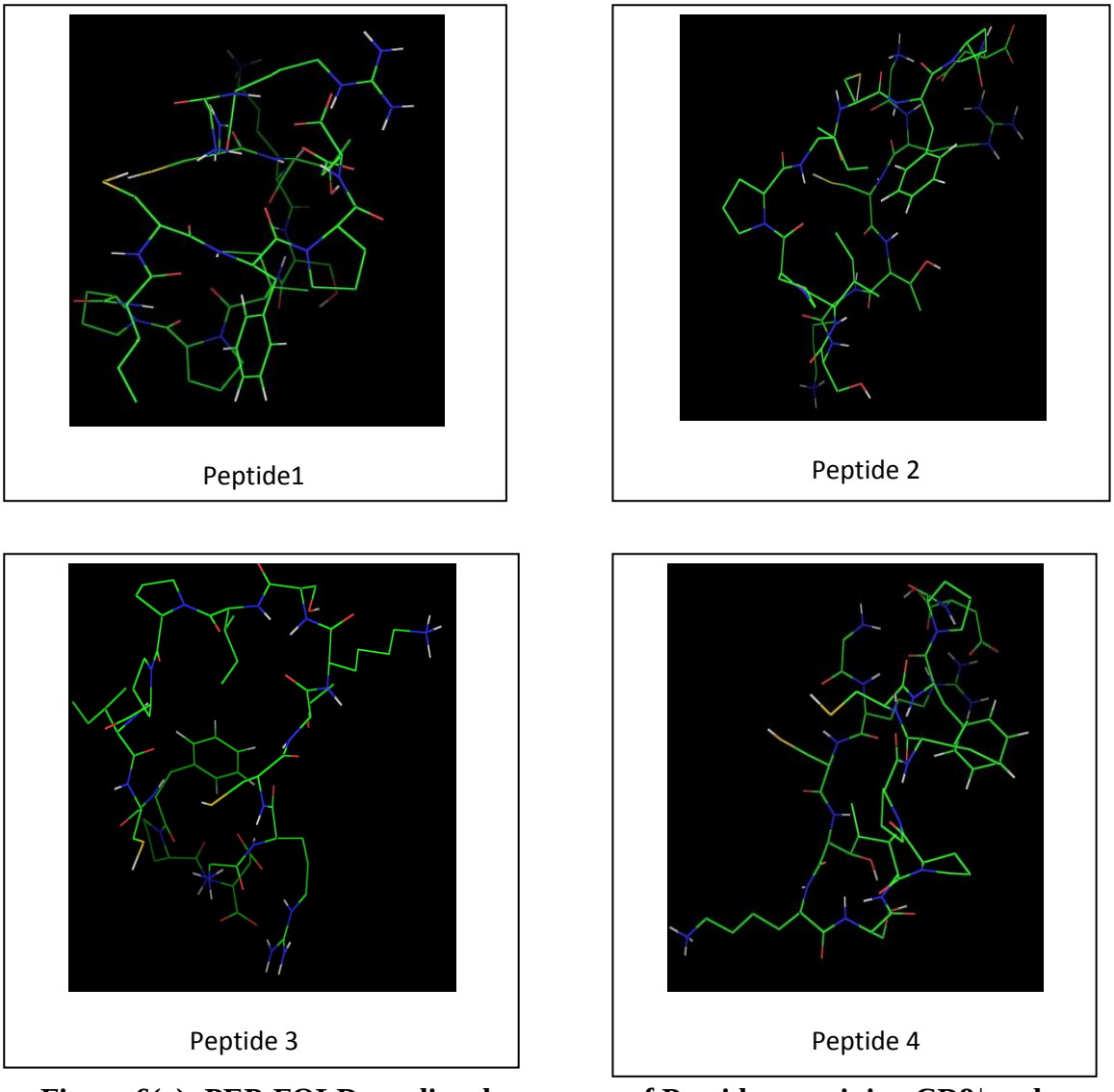
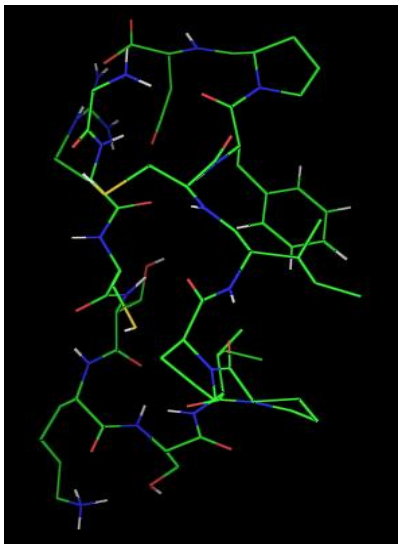
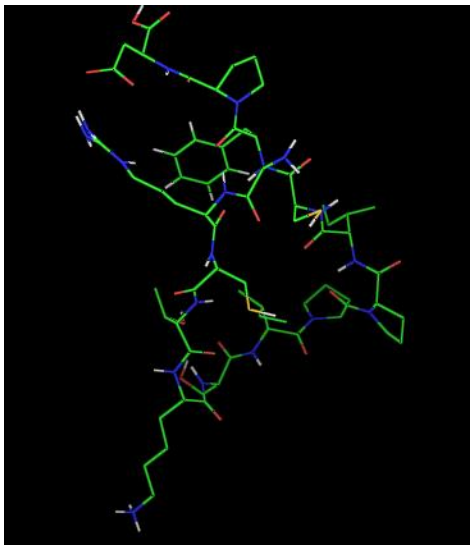


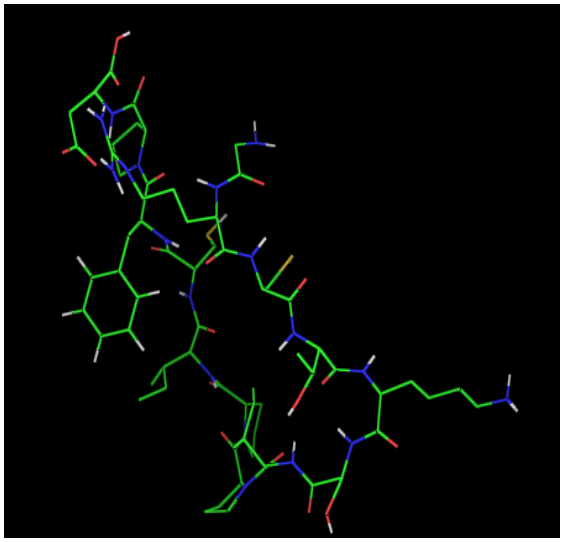
Figure 6(a): PEP-FOLD predicted structure of Peptide containing $CD8^+$ and $CD4^+$ specific T-cell epitopes (Peptide1, Peptide2, Peptide3 and Peptide4).



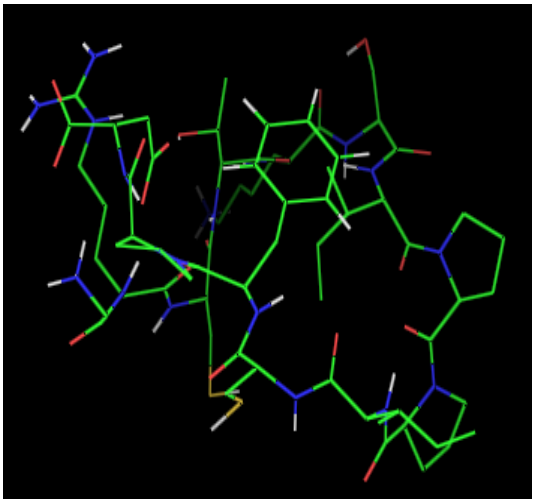
G1



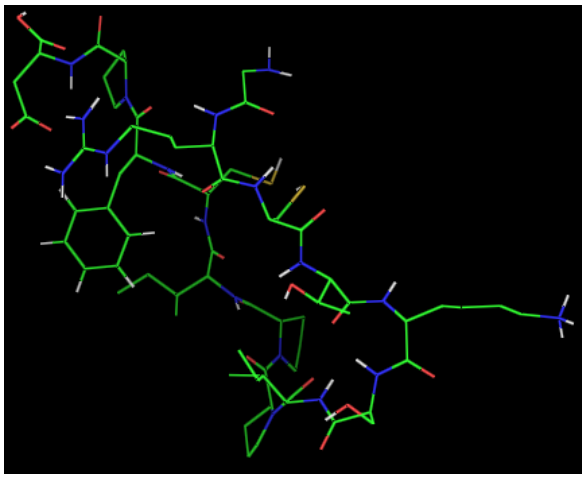
G2



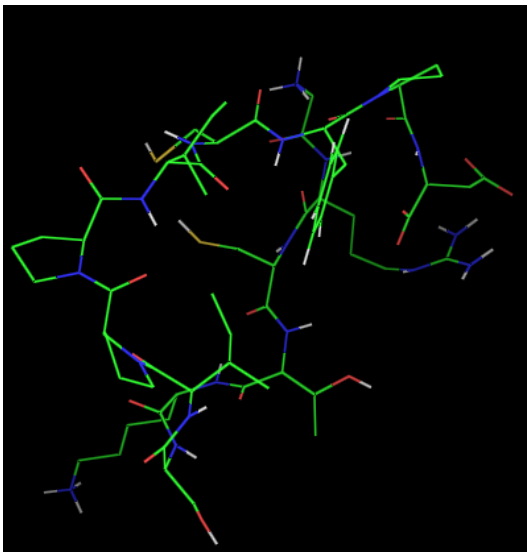
G3



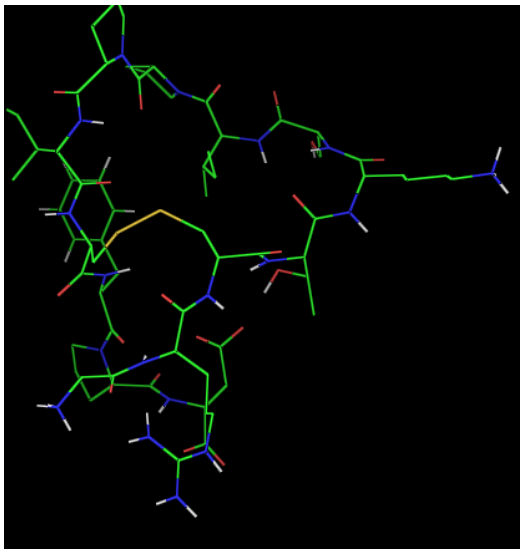
G4



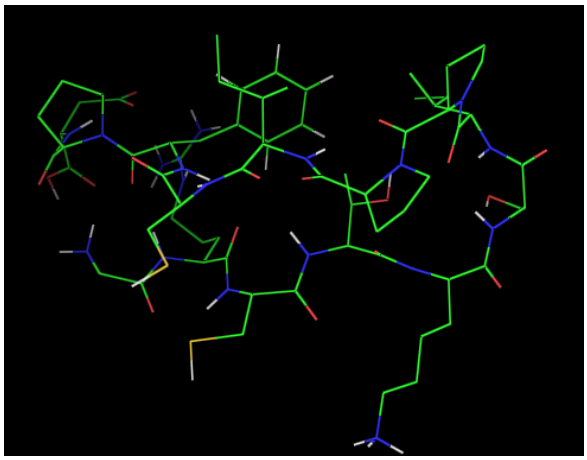
G5



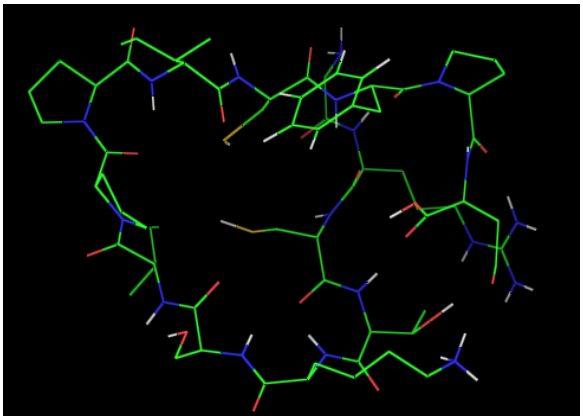
G6



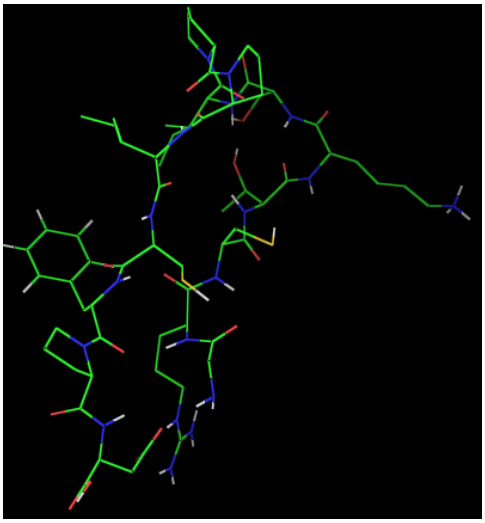
G7



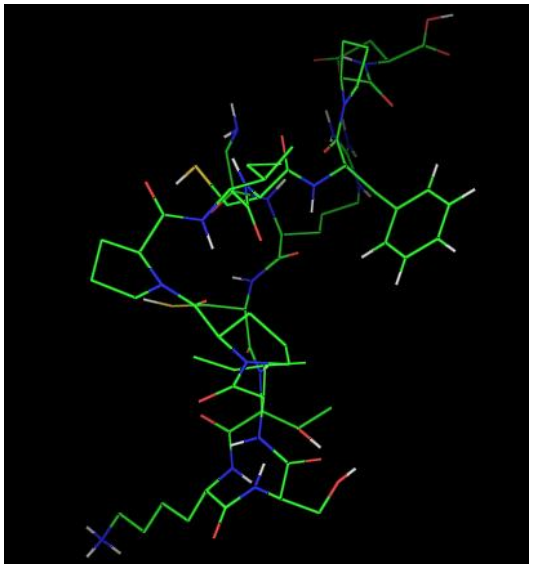
G8



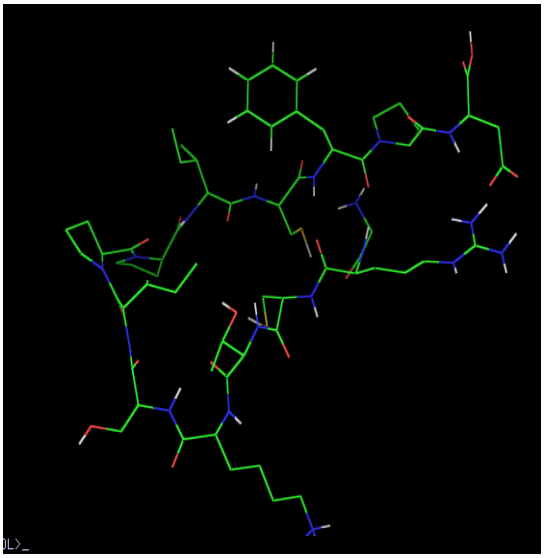
G9



G10



G11



G12

Figure 6(b) PEP-FOLD predicted structure G1 to G12 nanomer; Predicted Structure of G12 is a common predicted epitope for both $CD8^+$ and $CD4^+$ T cells.

5.6 Epitope Docking with Class I and II HLA molecules

The immunogenic potential of the predicted epitopes depends upon the fact that how efficiently it can bind to peptide binding pocket of the HLA molecule. The auto dock vina tool has been used to assess the binding affinity of the predicted epitope with HLA class I and II molecules. Two peptide structure HLA complex for Class I and II respectively were retrieved from PDB database for docking purpose. Peptides were extracted from the HLA complex and then redocked with corresponding HLA molecules, using auto dock vina. Twelve epitopes (G1to G12) as given in table 7 were docked with HLA Class I molecule while four peptide fragments (Table 6) containing the overlapping epitopes specific for both HLA Class I and Class II epitopes were docked with HLA Class II molecule.

The individual binding energies of different peptides after docking with class II HLA molecule were found to be comparable to the binding energy of the naturally bound peptides (Table 9).

Similarly the binding energies of the epitopes (G1to G12) which are specific for CD8+ T cell epitopes as given in Table 2 were docked with HLA Class I molecule. For HLA class I molecule, HLA-DQ2 structure was used for docking. The binding energies of the peptides capable of binding to HLA molecule were compared with the naturally bound peptide; given in Table 10.

Table 9: Binding Energies of peptide fragments for $CD4^+$ T cell with Class II HLA alleles

	Binding Energy (Kcal/mol)				
Class I HLA Alleles	Redocked Naturally Bound Ligand	Peptide1	Peptide2	Peptide3	Peptide4
B8	8.4	7.0	7.3	7.5	6.8

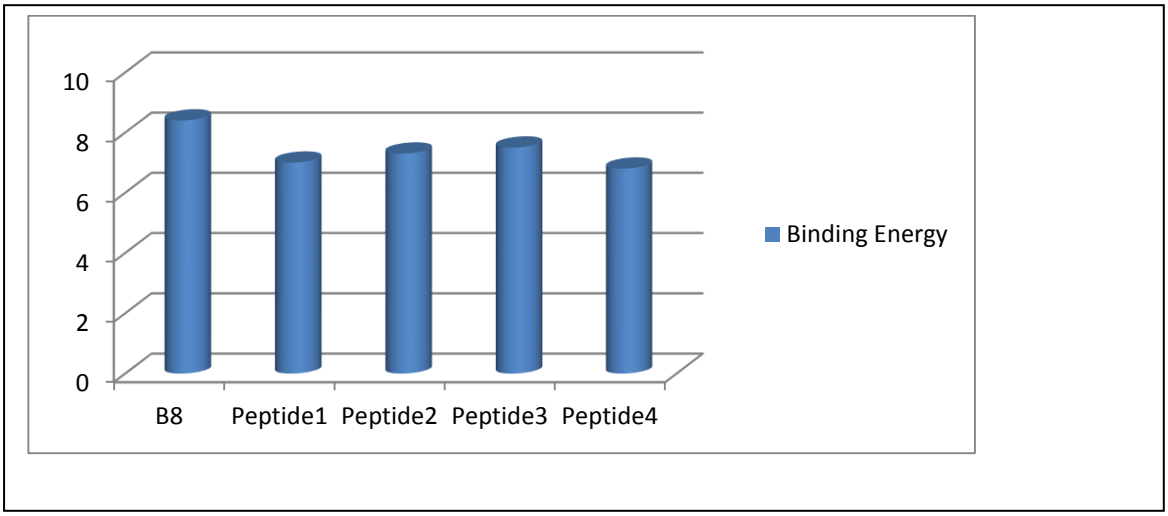


Figure 7: Binding energy of naturally bound peptide and Peptide1, Peptide2, Peptide3, Peptide4 obtained after docking to B8 class II HLA allele (Naturally bound Peptide).

Table 10: Binding Energies of epitopes for $CD8^+$ T cell with Class I HLA alleles

Epitopes	Binding Energy(Kcal/mol)
G1	7.7
G2	7.6
G3	7.4
G4	6.7
G5	7.6
G6	7.5
G7	8.7
G8	7.9
G9	7.7
G10	8.0
G11	7.7
G12	7.3
Natural bound peptide	8.7

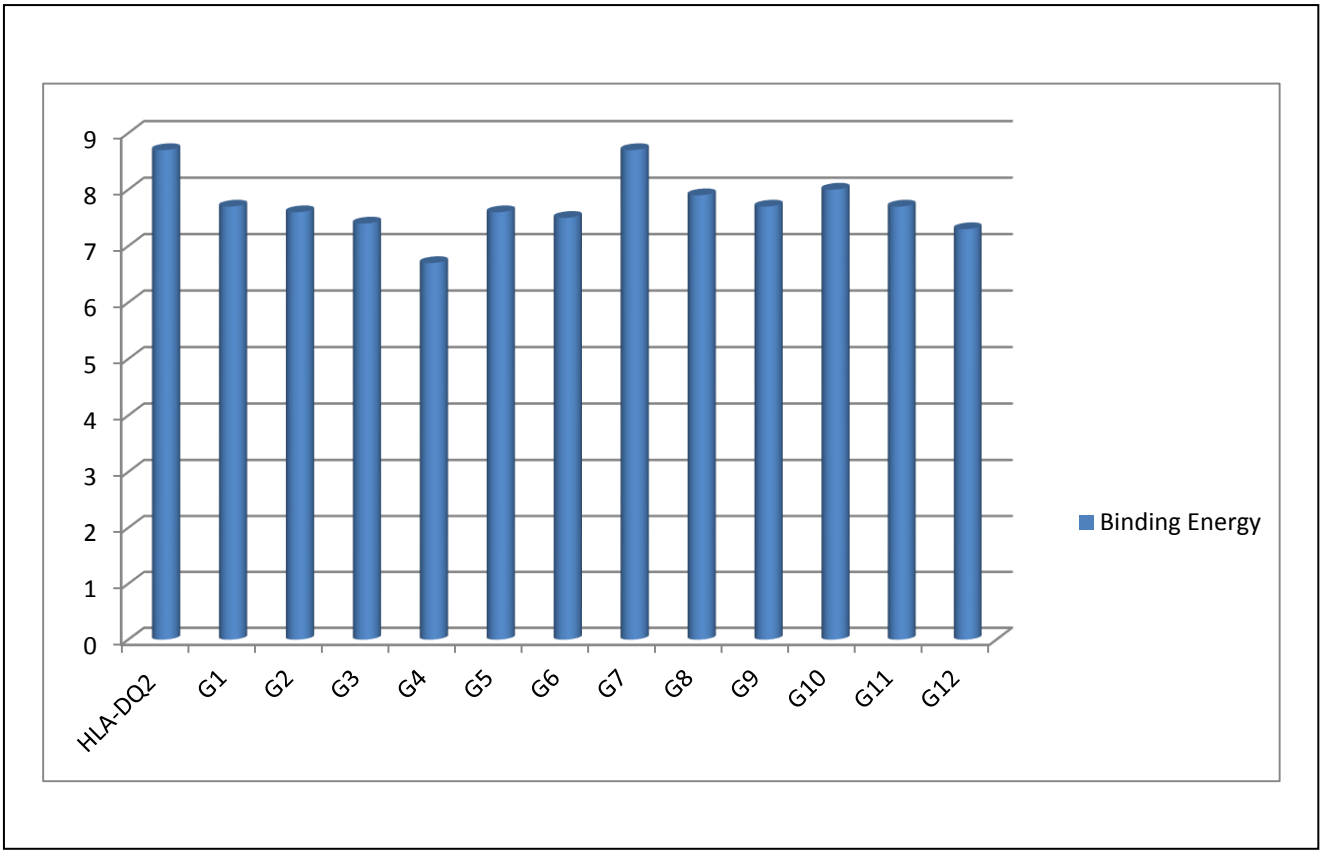
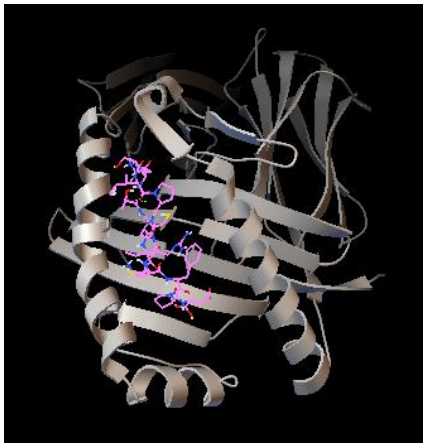


Figure 8: Binding energy of naturally bound peptides and G1 to G12 epitopes obtained after docking to HLA-DQ2 class I HLA allele (Naturally bound Peptide).

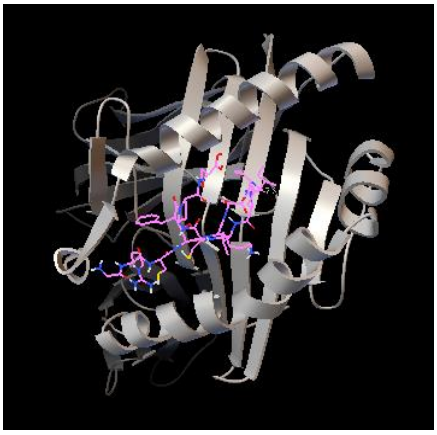
Molecular docking results demonstrate that NRKAIDFLL (G7) had same binding energy as compared to naturally bound peptide for Class I HLA alleles while peptide FHKEGAFLYDRLASTVIY (Peptide 3) had comparable binding energy to naturally bound peptide for Class II HLA alleles. Hence, these peptides will be strongly presented by HLA molecules present on antigen presenting cells to induce T cell response.

The docking poses of the four conserved peptides docked to the peptide binding pocket of HLA DQ2 are given in figure 9. Out of the twelve epitopes docked to the peptide binding pocket of HLA B8, docking poses of four epitopes are shown in figure 10 which had most comparable binding energy with natural bound peptide.

Docking Poses of HLA B8

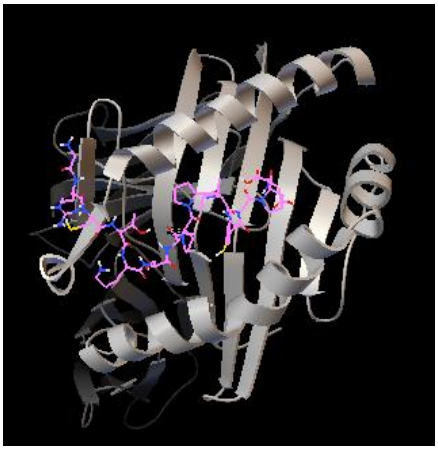


Peptide 1

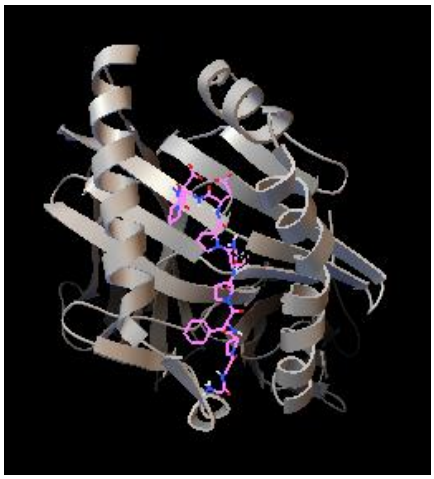


Peptide 2

Peptide 3



Peptide 4

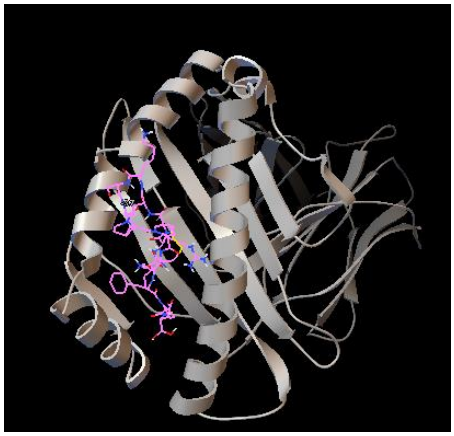


Natural bound peptide

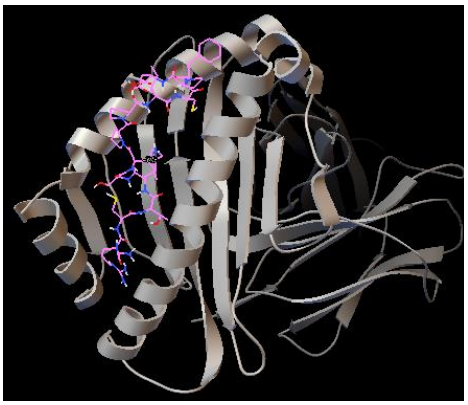
Figure 9: The peptide 1, peptide2, Peptide3 and Peptide4 are docked to the peptide
binding pocket of HLA B8.

As that of naturally bound ligand showing binding ability of Peptide1, peptide2, Peptide3 and Peptide4 towards HLA B8

Docking Poses of HLA DQ2

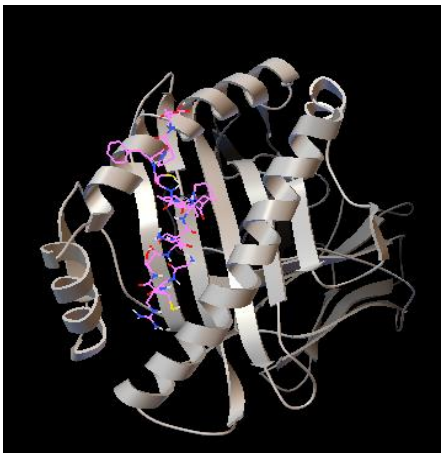


G7

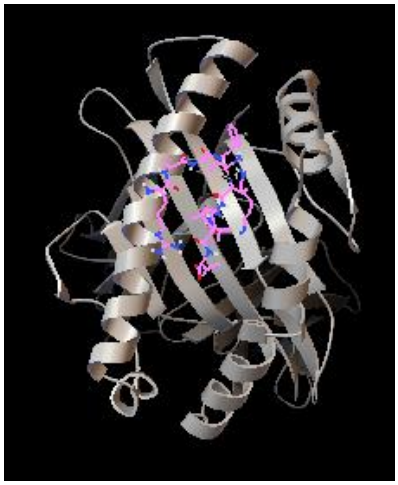


G8

G10



G12



Natural Bound peptide

Figure 10: The peptide G7, G8, G10 and G12 are docked at the site peptide binding pocket of HLA DQ2 whereas of naturally bound ligand is docked into epitope binding pocket showing non-binding ability of G7, G8, G10 and G12 towards HLA DQ2.

G12 is common for both CD8+ and CD4+ specific nonamer epitopes

One of the major arm of immune response is humoral immune response i.e. antibody mediated response. B cell produces antigen specific antibodies hence it will be interesting if the predicted peptide have the potential to generate humoral immune response. Linear B cell epitope prediction of Glycoprotein of Ebola virus was carried out by IEDB Kolaskar and Tongaonkar tool. Predicted B cell epitopes were compared with finally selected peptides containing T cell epitopes and four of the predicted immunogenic peptides were found to be common for both T and B cell epitope (Table 11)

Table 11: Predicted B-cell epitopes and peptides showing both T and B cell epitopes in GP protein of Ebola.

Peptide	**Peptides containing $CD4^+$ and $CD8^+$ T cell epitopes**	**Peptides containing both T and B cell epitopes**
FKRTSFFLWV	DRFKRTSFFLWV	**DRFKRTSFFLWV**
ILFQRTFSIPLG	ILFQRTFSIPL	**ILFQRTFSIPLG**
IRGFPRCRYVHVSGTGPCAGDFAFHKEGAFFLYDRLASTVIYRGTTF	FHKEGAFLYDRLASTVIY	**IRGFPRCRYVHVSGTGPCAGDFAFHKEGAFFLYDRLASTVIYRGTTF**
RKAIDFLLQRWGG	LRQLANETTQALQLFLRATTELRTFSILNRKAIDFLL	**LRQLANELRQLANETTQALQLFLRATTELRTFSILNRKAIDFLLQRWGG**

Chapter 6 DISCUSSION

Ebola virus disease is the causative agent of the Ebola virus disease known to produce a severe Hemorrhage fever in humans. Current treatment for Ebola is only symptomatic as no proven treatment is yet available. There are no licensed vaccines against Ebola, though two potential candidates are under clinical trials.

Immunoinformatics driven vaccine design research is emerging as a new way to lessen time in vaccine design process**.** It bypasses the exhaustive and expensive conventional vaccine design approach and uses the *in silico* vaccine design algorithms providing a practical alternative. Epitopes predicted by various bioinformatics tools have proven to be immunogenic both *in vivo* and *in vitro* assays. There have only been very few studies which have aimed at the prediction of epitopes in Ebola through *in silico* approach. Majority of them have used only one tool or two tools for epitope prediction. The prediction of HLA class II restricted T cell epitopes for the Ebola virus (EBOV) envelope glycoprotein (GP) has been done using IEDB (Ponomarenko *et al.,* 2014).

The current study considered 144 protein sequences of Ebola Glycoprotein which is the largest of its kind. Instead of only targeting the selected strains, we considered all the protein sequences of Ebola GP present since the first outbreak till the latest 2014 West African outbreak of Ebola. This led to consideration of all the four strains of Ebola known to infect human population. Application of six different algorithms (different tools) in our study was to ensure the immunogenic potential of the predicted epitopes. There are various factors which can invoke an immune response. NetCTL tools integrate prediction of peptide HLA class I binding proteasomal C terminal cleavage and TAP transport efficiency. HLA class I binding and proteasomal cleavage is performed using artificial neural networks. TAP transport efficiency is predicted using weight matrix. The scores from the three individual prediction methods are integrated as a weighted sum with a relative weight on peptide/HLA binding of 1(Larsen M.V. *et al.,* 2007).

Similarly SYFPEITHI contains a collection of HLA class I and class II ligands and peptide motifs of humans and other species. The prediction is based on motif matrices of published motifs and takes into consideration the amino acids in the anchor and auxiliary anchor positions, as well as other frequent amino acids.
BIMAS on the other hand predicts the epitopes based on the dissociation half life of the HLA complexes. Hence, using multiple tools for prediction of the same epitopes predicted by different algorithms strengthen the potential of putative epitopes for immunogenicity. Further the selection of common epitopes enriched peptide fragments for both Class I and Class II HLA may give rise to overall T-cell specific immune response. At the same time similarity between T-cell and B-cell linear epitopes may be helpful in choosing vaccine candidates capable of generating overall immune response.

Also to avoid any chance of auto immune response and eliminating the chances of tolerance by the predicted peptides BLAST analysis is of great importance. Thus BLAST screening eliminated the seven HLA class I and two class II specific epitopes to rule out any chance of auto immune response. Hence the current study reports four peptides which were generated containing overlapping $CD8^{+}$ and $CD4_{+}$ T cell epitopes. All these peptide fragments are 90% conserved.
The fragment **LRQLANETTQALQLFLRATTELRTFSILNRKAIDFLL**, out of the final four peptide fragments, contained common epitope for both CD8+ and CD4+ T cell cells as well as for B cell. This peptide fragment also contained maximum number of predicted epitopes both for CD8+ and CD4+ T cell. After studying the crystal structure of GP protein, it was found that this peptide lies in the heptad repeat regions HR1 and HR2 of GP2. HR1 region of GP2 protein is found to be conserved among all members of Filovirus and thus this peptide fragment can prove a very interesting candidate for vaccine design against Ebola and Marburg Virus of the Filoviridae family (Lee *et al.*, 2008).

Also the peptide fragment **FHKEGAFFLYDRLASTVIY** which contained overlapping $CD8^{+}$ and $CD4^{+}$ T cell epitopes was found in the head region of GP1 part of the GP between residues 54-201. This region mediates virus cell surface

attachment to host cells. Hence this peptide also can prove very beneficial in generation of therapeutics against Ebola (Lee *et al.,* 2008).

Another additive advantage of computational epitope prediction is that the epitope predictions are made specific to HLA allele. Thus it enabled us to select those putative epitopes which are recognized by large number of HLA class I and II allele. HLA alleles are highly polymorphic and diversity of alleles were found in the global population. Different individuals of the worldwide population may respond in different way against particular antigen. The IEDB database enables predicting the possible world population capable of responding to particular immunogen. The database contains variety of HLA class I and II alleles frequently found in different world population of geographical areas. In our study IEDB population coverage analysis tool was carried out for 16 different geographical areas distributed globally (Huynh-Hoa Bui *et al.,* 2006). Interestingly, peptide fragments containing $CD8^+$ Tcell epitopes revealed 100% population coverage in Europe, West Africa, East Africa, Central Africa, North America and Central America. Peptide fragments containing $CD4^+$ T cell epitopes revealed 99% in case of Europe, East and South East Asia. These results show that these peptide fragments are capable of inducing an immunogenic response among individuals belonging to different regions of the world population.

Molecular docking is extensively used for computer aided drug design but in recent years it is also being used for peptide HLA interaction study. All 12 predicted nonamer peptides which contained CD8+ T cell epitopes and four peptides fragments containing overlapping $CD8^+$ and $CD4^+$ T cell epitopes were considered for docking studies because core binding region of peptide to HLA molecule is nonamer. All the bound peptides showed comparable binding energies with the natural bound peptide. Similar kind of studies has been reported for nucleoprotein of H5N1 virus where they compare the binding energy with predicted epitope and bound peptides for only class I HLA molecules (Hou *et al.,* 2012).

Currently there are two potential vaccines candidates which are undergoing phase II and phase III human clinical trials in Africa. ChAd3-ZEBOV is developed by GlaxoSmithKline (GSK), in collaboration with the US National Institute of Allergy and Infectious Diseases (NIAID) and rVSV-ZEBOV, developed by NewLink Genetics and Merck Vaccines USA, in collaboration with the Public Health Agency of Canada. Both the vaccine are vectored vaccines in which the Ebola virus glycoprotein is vectored in replication-incompetent chimpanzee adenovirus 3 (cAd3) or a replication-competent vesicular stomatitis virus (VSV). But the immunogenicity and reactogenicty of both are dose dependent as the higher does required to generate an immune response also produced adverse effects in 70% candidate population (Bausch *et al.,* 2015). One of the main reasons for lack of any therapeutics is that the complete molecular biology and pathogenesis of Ebola virus including virulence factors and host responses still remains unclear. This is due, in part, to the difficulty in obtaining samples and studying the disease in the relatively remote areas in which the outbreaks occur. In addition, a high degree of biohazard containment is required for its laboratory studies and clinical analysis. Hence use of conserved peptides containing epitopes would be an ideal candidate for preparing a vaccine in such case.

Chapter 8. REFERENCES

- Brauburger Kristina, Boehmann Yannik, Tsuda Yoshimi *et al*., Analysis of the highly diverse gene borders in Ebola virus reveals a distinct mechanism of transcriptional regulation, Journal of Virology,2014 , Vol 88, 12558 –12571
- Becquart Pierre, Mahlako Tanel, Dieudonne´ Nkoghe *et al*., Dentification of Continuous Human B-Cell Epitopes in the VP35, VP40, Nucleoprotein and Glycoprotein of Ebola Virus. PLoS ONE,2014,Vol9(6),1-8
- Changulaa Katendi, Yoshidac Reiko, Noyoric Osamu *et al*., Mapping of conserved and species-specific antibody epitopes on the Ebola virus nucleoprotein,Journal of Virology,2013, 83–90
- Choi Woo Young, Hong Kee-Jong, Hong Joo Eun, Progress of vaccine and drug development for Ebola preparedness, Clinical and Experimental Vaccine Research,2015,Vol4(1),11-16

- Dawson Deborah, Ozgur Mehmet, Sari Kemal, Ramifications of HLA Class I Polymorphism and Population Genetics for Vaccine Development, Genetic Epidemiology,2001,Vol20, 87–106
- Dorival Isabel, Wu Weining, Dowall Stuart *et al*., Elucidation of the Ebola virus VP24 cellular interactome and disruption of virus biology through targeted inhibition of host cell protein function, J. Proteome Res, 2014,1-52
- Feldmann Heinz, Geisbert Thomas, Ebola hemorrhagic fever, Lancet 2011,Vol. 377,849–862
- Gaya Amarasinghe, Christopher Basler, Evasion of Interferon Responses by Ebola and Marburg Viruses, Journal of Interferon & Cytokine Research, 2009, Vol.29(9),511–52

- Geisbert W Thomas. and Feldmann Heinz, Recombinant Vesicular Stomatitis Virus–Based Vaccines Against Ebola and Marburg Virus Infections, The Journal of Infectious Diseases,2011,Vol 3,1075-1080
- Gupta Manisha, Mahanty Siddhartha, Greer Patricia *et.al*., Persistent Infection with Ebola Virus under Conditions of Partial Immunity, JOURNAL OF VIROLOGY, 2004, Vol. 78, 958–967
- Gulukota Kamalakar, Charles DeLisi, HLA allele selection for designing peptide vaccines,Analysis: Biomolecular Engineering,1996,Vol 13,81-86
- Korber Bette, LaBute Montiago, Yusim Karina, PLoS Comput Biol,2006, Vol 2(6), 0484- 0492
- Lee Jeffrey, Fusco Marnie, Oswald Wendelien, Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor, Nature July 2008 Vol. 454,177-180
- Martinez Osvaldo, Leung Lawrence, Basler Christopher , The role of antigen-presenting cells in filoviral hemorrhagic fever: gaps in current knowledge Antiviral Research, March 2012, Vol. 93(3),Pages 416–428
- Mohan Gopi , Wenfang Li, Ye Ling , Compans Richard, Antigenic Subversion: A Novel Mechanism of Host Immune Evasion by Ebola Virus, PLoS Pathogen, December 2012,Vol.8(12),1-14
- Morten Nielsen, Ole Lund, Soren Buus and Claus Lundegaard, HLA Class II epitope predictive algorithms, Immunology,2010,Vol 130, 319–328
- Parker K.C., Bednarek M.A. and Coligan J.E ,Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains‖ J. Immunol., 1994,w3152, 163-175
- Perez Beatriz, Volchkova Valentina, Dolnik Olga, Shed GP of Ebola Virus Triggers Immune Activation and Increased Vascular Permeability, PLoS Pathog,2014, 10(11): e1004509. doi:10.1371/journal.ppat.1004509
- Rammensee, Bachmann, Stevanovic, HLA ligands and peptide motifs, Landes Bioscience, 1997

- Rosa Daniela, Ribeiro Pereira Susan, Cunha-Neto Edecio, $CD4^+$ T Cell Epitope Discovery and Rational Vaccine Design, Arch. Immunol. Ther. Exp,2010,Vol 58, 121–130
- Reid Patrick, Shurtleff Amy, Costantino Julie, HSPA5 is an essential host factor for Ebola virus infection, Antiviral Research 2014, Vol.109, 171–174
- Reche Pedro, Caldas Enrique, Flower Darren, Peptide-Based Immunotherapeutics and Vaccines, Journal of Immunology Research,2014, Volume 2014 Article ID 256784
- Sullivan Nancy, Yang Zhi-Yong, Nabel Gary, Ebola Virus Pathogenesis: Implications for Vaccines and Therapies, JOURNAL OF VIROLOGY, Sept. 2003, 9733–9737
- Singh H. and Raghava G. P. S, ProPred: prediction of HLA-DR binding site‖ BIOINFORMATICS APPLICATIONS NOTE, 2001, Vol. 17(12), 1236–1237
- Sullivan Nancy, Yang Zhi-Yong, Nabel Gary, Development of a preventive vaccine for Ebola virus infection in primates, Nature November 2000,Vol 408,605-609
- Sundar Krishnan , Boesen Agnieszka, Coico Richard, Computational prediction and identification of HLA-A2.1-specific Ebola virus CTL epitopes, Elsevier Virology 2007, Vol 360 ,257–263
- Thi Le Thuy, Sarmiento Maria, Calero Romel *et al*., Immunoinformatics study on highly expressed Mycobacterium tuberculosis genes during infection, Tuberculosis 2014,475-481
- Volchkov Viktor, Volchkova Valentina, Slenczka Werner *et al*., Release of Viral Glycoproteins during Ebola Virus Infection, Virology Academy Press, 1998,110-119
- WHO Fact sheet N°103 , Ebola virus disease, Updated September 2014
- Yen Benjamin , Mulder Lubbertus, Martinez Osvaldo *et al*., Molecular Basis for Ebola Virus VP35 Suppression of Human Dendritic Cell Maturation, Journal of Virology,2014,1-42

- Zhang Qing, Wang Peng, Kim Yohan *et al*., Immune epitope database analysis resource, Nucleic Acids Research, 2008, 2008, Vol. 36,514-516