4.1 Methodology for studying disease process caused by Hepatitis C virus

4.1.1 Uncharacterized protein NS3a sequence

The sequence of the uncharacterized NS3 protein, subtype 3a was obtained from the UniProt database which was further used for the epitope prediction.

>trf|D6BQ42|D6BQ42_9HEPC NS3 (Fragment) OS=Hepatitis C virus subtype 3a GN=NS3 PE=4 SV=1

4.1.2 Epitope prediction

The epitopes (9-mer peptides) from the sequence NS3a protein of HCV were predicted for all the Human Leukocyte Antigen (HLA) alleles for both the MHC Class I and MHC Class II molecules. The prediction for promiscuous MHC binders was carried out via ProPred-I (Singh et al., 2001) and ProPred (Singh and Raghava, 2003) for MHC Class I and MHC Class II respectively.

4.1.3 Virtual Screening

In this step, we screened the promiscuous MHC binders separately for both the MHC Class molecules. From the result table obtained for epitope prediction, each epitope were counted and arranged in the decreasing order on the basis of the number of predictions obtained for multiple alleles as binders. Further conservancy analysis was done to screen peptides which proved to be binding with both class of MHC molecule. This will prove to be a better vaccine as epitope binding to only one of the MHC molecule will not be so effective for producing immunity against such viral agent (Parida et al., 2007) We obtained six peptide sequences (epitopes) after virtual screening which were the binders to both MHC Class I and MHC Class II molecules. These are, (LLGTIVTSL; FLGTTVGGV; VVTGEVQVL; LVTRDADV1; VLSTATQTF; IPARRRGDS).

4.1.4 Structure Prediction

The structure prediction of above six predicted epitopes was carried out by using the CPHmodel-3.2 Server, an online three dimensional structure prediction sever which is based on automated homology modeling.
4.1.5 Energy simulation

The predicted structures of the above peptide sequences was energy optimized using the tool GROMACS (Groningen Machine for Chemical Simulation) (Blau and Grubmuller, 2013), which is itself a matter of intense study works well on the LINUX environment. It uses the five dynamic steps these are, pdb2gmx; editconf; genbox; grompp and mdrun. The energy optimization is necessary in order to decrease the electronic repulsion and steric hindrance and for the favorable local interactions between the amino acid side chains so as to get a stable native and functional conformation.

4.1.6 Validation

After checking the stability of the peptide molecules, these peptides were validated by using the online validation server RAMPAGE based on the concept of Ramachandran Plot. The structures were validated to confirm its stable secondary structure conformation. This was validated by analyzing the number of amino acid residues in favoured region, allowed region and disallowed region. It was found that the peptide sequence VLSTATQTF showed the amino acid residues in the favored region with 100 per cent. And at second position, the peptide sequence LLGTIVTSL showed the amino acid residues in the favored region with 83.3 per cent. Other peptide sequences showed either 75 per cent or less than that. This again can be a step of screening but instead we used all six epitopes structure for docking with MHC Class I and MHC Class II molecule.

4.1.7 MHC Class I and MHC Class II structure

3D structure of MHC Class I and MHC Class II(PDB ID: 3MGR; PDB ID: 1DLH) respectively of Human origin were taken from PDB database.

3MGR is the crystal structure of Human MHC Class I HLA-A2 molecule complexed with HCV and 1DLH is the crystal structure of Human MHC Class II (HLA-DR1, Alpha chain) and (HLA-DR1, Beta chain) complexed with enterotoxin type B precursor of Influenza virus. These complexed ligands were then removed using three dimensional molecular visualization tools pymol to obtain the pure crystal structure of MHC molecules.

4.1.8 Docking studies:

We performed the docking studies of all the six peptide structures separately each with MHC Class I (PDB ID: 3MGR) and MHC Class II (PDB ID: 1DLH) molecule using an online
docking tool named as PatchDock (Schneidman-Duhovny et al., 2005). Docking score and atomic coordinate energy (ACE) value were considered to find out the best epitopic peptide. In this study, it was found that the docking of the peptide sequence (LLGTIVTSL) with MHC Class I (PDB ID: 3MGR) showed the top most score of 8896 and the least ACE value of -255.14 KJ/Cal as compared to other peptide sequences. In second case, the docking of peptide sequence (LLGTIVTSL) with the MHC Class II (PDB ID: 1DLH) again showed the highest score of 9012 with least ACE value of -300.73 KJ/Cal, as compared to other peptide sequence.

4.2 Methodology for studying literature based analysis of interaction between all possible pairs of viral-human proteins

4.2.1 Extraction of Protein-Protein interactions Data
Extraction of Protein-protein interactions between different proteins present in Hepatitis C Virus and proteins present in human. To extract protein-protein interaction data “HCVpro” database is utilized. HCVpro is an online database for the protein-protein interaction between Hepatitis C Viral proteins and human proteins. Highly specific protein interactions extracted in a strain specific manner. HCVpro provide absolute interaction data containing the PubMed identifiers for each interaction.

4.2.2 Preparation of protein specific interaction list
The enormous raw form of protein-protein interactions was extracted from the “HCVpro” database and processed manually in a protein specific manner. Interactions of each of the Hepatitis C proteins are listed separately.

4.2.3 Analysis of interactions maps
Detailed analysis of protein-protein interaction maps was done by studying different interactions between Hepatitis C proteins and human proteins. A specific protein-protein interaction is selected on the basis of its significance in the pathway leading to replication of Hepatitis C genome i.e. Interaction between Hepatitis C Nonstructural protein 5A (NS5A) with SH3 domain of Fyn tyrosine protein kinase.

4.2.4 Analysis of protein-protein interaction
Selected protein-protein interaction is studied to find out type of interaction between NS5A and Fyn tyrosine protein kinase by mining of PubMed research articles. This interaction
provided insights that NS5A can be targeted so as to develop potential vaccine against Hepatitis C Virus.

4.2.5 Prediction of T-cell Epitopes for target Protein
Protein sequence in fasta format of NS5A is used to predict MHC Class I and MHC Class II binding peptides by using “Propred” for MHC-II (Singh and Raghava, 2003) and “Propred-I” for MHC I (Singh and Raghava, 2001). Propred is an online server for the prediction of MHC II binding peptides for 51 alleles, whereas Propred-I is an online server for the prediction of MHC I binding peptides for 47 alleles.

4.2.6 Virtual Screening
In this step, we screened the promiscuous MHC binders separately for both the class of MHC molecules. Predicted T-cell epitopes were screened on the basis of total average score and number of alleles. We obtained 6 peptide sequences (epitopes) after virtual screening which were the binders to both MHC Class I and MHC Class II molecule. These were, (DEITFMVGL; WRGDGVMST; VVILDSFEP; MVGLNSYRI; DVSVLTAML; TRCSCGATI) (Kaushik et al., 2013b).

4.2.7 Tertiary structure prediction for epitopes
Structure prediction is done by “PepFold” which is an online peptide structure prediction server. It uses four consecutive residues and predicts their conformations using greedy algorithm coupled with coarse grained force field. Only one conformation (model) is selected for each peptide on the basis of lowest sOPEP (OPEP-Optimized Potential for Efficient Structure Prediction) energy.

4.2.8 Docking using PatchDock and FireDock
Each of the predicted structures were docked with MHC Class I and Class II molecules using PatchDock and FireDock. PatchDock provides best binding peptide using the best binding score for each peptide. FireDock refines the PatchDock result and gives the best binding peptide using global energy of each docking (Schneidman-Duhovny D., 2005). Best epitopes having the lowest global energy and binding score were selected as potential high affinity binding epitope.
4.2.9 Calculation of Isoelectric point (pI value) and half-life of dissociation of selected epitopes
Isoelectric point and molecular weight of the epitopes was calculated using ExPASy tool online. Epitopes with lowest pI value were selected.

4.3 Methodology for predicting epitopic peptides by using existing tools

4.3.1 Prediction of T-cell Epitopes for target proteins
The HCV NS3 and NS5 protein of 3a and 3b genotype were retrieved from UniProt protein sequence database using ID: Q81258 for 3a genotype and ID: Q81487 for 3b genotype. For T cell epitope prediction five tools were used for better confirmation of good epitope. IEDB, RankPEP, HLA peptide motif, PREDEP and SYFPEITHI are the tool that were used and the top 20 epitope was selected without selecting repeated epitope of different HLA allele.

4.3.2 Virtual Screening
Screening of epitope was first done for each protein; epitopes that were common for all tools were selected. After that selected epitope were compared between 3a and 3b genotype to select the conserved epitope. From that conserved epitope only 11 epitope were selected by observing the properties i.e., flexibility, antigenicity, hydrophilicity and beta turn. The most promising thing which is observed in selection is that if any epitope satisfy one property completely then it may happen that other properties are not well satisfied by that epitope. So on the basis of property evaluation; epitopes are selected by giving more weightage to hydrophilicity and antigenicity.

4.3.3 Tertiary structure prediction for epitopes
Structure prediction is done by “PepFold” which is an online peptide structure prediction server. It uses four consecutive residues and predicts their conformations using greedy algorithm coupled with coarse grained force field. Only one conformation (model) is selected for each peptide on the basis of lowest sOPEP (OPEP-Optimized Potential for Efficient Structure Prediction) energy.

4.3.4 Docking Studies
Each of the predicted structure were docked with MHC Class I and class II molecules using PatchDock and FireDock. PatchDock provides best binding peptide using the best binding
score for each peptide. FireDock refine the PatchDock result and gives the best binding peptide using Global energy of each docking (Bartenschlager et al., 1993). Best epitopes having the lowest global energy and binding score were selected as potential high affinity binding epitope.

4.4 Methodology for performing docking and simulation of these screened peptides

4.4.1 T-cell Epitopes prediction of target proteins

The NS3 as well as NS5 protein of HCV with 3a and 3b genotype were fetched from UniProt database. The 3a genotype was retrieved using ID: Q81258 whereas 3b genotype from ID: Q81487. Five tools were worn out for excelling confirmation of the good epitope in T cell epitope prediction, are RankPEP, SYFPEITHI, IEDB, PREDEP, and HLA peptide motif. With the help of these tools, top 20 epitopes were recruited without repetition for different HLA allele.

4.4.2 Virtual Screening

Screening of epitope was done for respective proteins; epitopes that were common for all tools were selected. After exploration, selected epitope were compared between 3a and 3b genotype to ascertain conserved epitope. The conserved epitopes undergo selection, only 11 epitope were recruited by considering properties like antigenicity, flexibility, beta-turn, and hydrophilicity. The most assuring fact that was discovered while the selection was if any epitope accomplish one property then it is entirely not necessary that it will fulfil other properties too. With respect to property evaluation, the epitopes are recruited on a weightage of hydrophilicity as well as antigenicity.

4.4.3 Prediction of Tertiary structure for epitopes

PepFold, an online server for structure prediction was used for predicting peptide structure. Four constant residues were used in this process, whereas their conformations were predicted using a greedy algorithm clasp with the unrefined force field. Only one conformation (model) is selected for each peptide on the basis of lowest sOPEP (OPEP-Optimized Potential for Efficient Structure Prediction) energy.
4.4.4 Docking Studies

The anticipated structure was docked through PatchDock and FireDock with molecules of MHC Class I and II. PatchDock provides best binding peptide using the best binding score for each peptide. FireDock refines the PatchDock prediction and gives the best binding peptide using Global energy of each docking (Bartenschlager et al., 1993). Best epitopes having the lowest global energy and binding score were selected as potential high affinity binding epitopes.

4.4.5 Molecular Dynamics of Simulation of Allele epitope Complex

This type of simulation was carried out using the NAMD (Nanoscale Molecular Dynamics Program; v2.7) graphical interface module (Phillips et al., 2005) and incorporated visual molecular dynamics (VMD 1.9.2) (Humphrey et al., 1996).

4.5 Methodology for developing novel technique to enhance the accuracy and efficiency as given by the existing models

4.5.1 Collection and description of data

We have collected epitope sequences data from the download section of “Lbtope: B-cell linear epitope prediction server” (http://www.imtech.res.in/raghava/lbtope/index.php) that corresponds to 1042 B-cell epitopes and 1795 non epitopes. The epitope sequence length range from minimum 6 to maximum 48 for positive epitope dataset and minimum 6 to maximum 33 for negative epitopes dataset (i.e., non-epitope dataset). For the purpose of 10 fold cross validation (Arlot and Celisse, 2010) 10 we have prepared 10 different combinations of training and test datasets by picking 90% and 10% data randomly and repeating this step 10 times following the work of Singh et al (2013) (Singh et al., 2013).

4.5.2 Computational formalism

All the computational works were carried out by mostly writing MATLAB (R2014b) codes and occasionally using standard MATLAB library functions. For prediction of epitopes and non-epitopes Elman back propagation network (Lin and Hong, 2011) was employed as classifier. Model performance parameters, such as, average accuracy, specificity, sensitivity and Matthew’s correlation coefficient (MCC) for training as well as testing the data, were calculated over the 10 datasets.
4.5.3 Selection of amino acid features

To overcome the challenge of making an epitope sequence (actually a string) ready to be utilized as numerical input in a computational model by preserving maximum biological information conveyed by it, in the pre-processing step, six physicochemical properties were selected for each of the amino acids (Table 6). These properties were hydrophobicity (Tanford, 1978), volumes of side chains of amino acids (Radzicka and Wolfenden, 1988), polarity (Ellington and Cherry, 2001), polarizability (Krishtal et al., 2009), solvent-accessible surface area (Connolly, 1983) and net charge index of side chains of amino acids (Janin et al., 1978). Selection of these properties was inspired by the work of Huang et al., (2010) to identify protein-protein interaction (Xia et al., 2010). Apart from considering the amino acids physicochemical properties, percentage content of individual amino acids along with dipeptide types whichever is present within the epitope-sequence were also included as illustrated in Figure 8.

**Table 6:** Six physicochemical properties are used to represent each amino acid in to numerical form (Xia et al., 2010).

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Hydrophobicity</th>
<th>Volume of side chains</th>
<th>Polarity</th>
<th>Polarizability</th>
<th>Solvent accessible surface area</th>
<th>Net charge index of side chains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala (A)</td>
<td>0.62</td>
<td>27.5</td>
<td>8.1</td>
<td>0.046</td>
<td>1.181</td>
<td>0.0072</td>
</tr>
<tr>
<td>Cys (C)</td>
<td>0.29</td>
<td>44.6</td>
<td>5.5</td>
<td>0.128</td>
<td>1.461</td>
<td>-0.0366</td>
</tr>
<tr>
<td>Asp (D)</td>
<td>-0.9</td>
<td>40</td>
<td>13</td>
<td>0.105</td>
<td>1.587</td>
<td>-0.0238</td>
</tr>
<tr>
<td>Glu (E)</td>
<td>-0.74</td>
<td>62</td>
<td>12.3</td>
<td>0.151</td>
<td>1.862</td>
<td>0.0068</td>
</tr>
<tr>
<td>Phe (E)</td>
<td>1.19</td>
<td>115.5</td>
<td>5.2</td>
<td>0.29</td>
<td>2.228</td>
<td>0.0376</td>
</tr>
<tr>
<td>Gly (G)</td>
<td>0.48</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0.881</td>
<td>0.1791</td>
</tr>
<tr>
<td>His (H)</td>
<td>-0.4</td>
<td>79</td>
<td>10.4</td>
<td>0.23</td>
<td>2.025</td>
<td>-0.0107</td>
</tr>
<tr>
<td>Ile (I)</td>
<td>1.38</td>
<td>93.5</td>
<td>5.2</td>
<td>0.186</td>
<td>1.81</td>
<td>0.0216</td>
</tr>
<tr>
<td>Lys (K)</td>
<td>-1.5</td>
<td>100</td>
<td>11.3</td>
<td>0.219</td>
<td>2.258</td>
<td>0.0177</td>
</tr>
<tr>
<td>Leu (L)</td>
<td>1.06</td>
<td>93.5</td>
<td>4.9</td>
<td>0.186</td>
<td>1.931</td>
<td>0.0517</td>
</tr>
<tr>
<td>Met (M)</td>
<td>0.64</td>
<td>94.1</td>
<td>5.7</td>
<td>0.221</td>
<td>2.034</td>
<td>0.0027</td>
</tr>
<tr>
<td>Asn (N)</td>
<td>-0.78</td>
<td>58.7</td>
<td>11.6</td>
<td>0.134</td>
<td>1.655</td>
<td>0.0054</td>
</tr>
<tr>
<td>Pro (P)</td>
<td>0.12</td>
<td>41.9</td>
<td>8</td>
<td>0.131</td>
<td>1.468</td>
<td>0.2395</td>
</tr>
<tr>
<td>Gln (Q)</td>
<td>-0.85</td>
<td>80.7</td>
<td>10.5</td>
<td>0.18</td>
<td>1.932</td>
<td>0.0492</td>
</tr>
<tr>
<td>Arg (R)</td>
<td>-2.53</td>
<td>105</td>
<td>10.5</td>
<td>0.291</td>
<td>2.56</td>
<td>0.0436</td>
</tr>
<tr>
<td>Ser (S)</td>
<td>-0.18</td>
<td>29.3</td>
<td>9.2</td>
<td>0.062</td>
<td>1.298</td>
<td>0.0046</td>
</tr>
<tr>
<td>Thr (T)</td>
<td>-0.05</td>
<td>51.3</td>
<td>8.6</td>
<td>0.108</td>
<td>1.525</td>
<td>0.0034</td>
</tr>
<tr>
<td>Val (V)</td>
<td>1.08</td>
<td>71.5</td>
<td>5.9</td>
<td>0.14</td>
<td>1.645</td>
<td>0.057</td>
</tr>
<tr>
<td>Trp (W)</td>
<td>0.81</td>
<td>145.4</td>
<td>5.4</td>
<td>0.409</td>
<td>2.663</td>
<td>0.038</td>
</tr>
<tr>
<td>Tyr (Y)</td>
<td>0.26</td>
<td>117.3</td>
<td>6.2</td>
<td>0.298</td>
<td>2.368</td>
<td>0.0236</td>
</tr>
</tbody>
</table>
Figure 8 Conversion of epitope sequences into numerical form

[A. Selection of amino acid one by one from epitope; B. Identified the position of selected amino acid from AA index; C. Selection of physicochemical property based on the position of selected amino acid followed by concatenation of amino acid and dipeptide composition calculated by considering individual epitope at a time].

4.5.4 Normalization of physicochemical properties

All the six physicochemical properties of amino acids were normalized by using one of the most cited, robust and highly efficient ‘tanh estimators’ (Hampel, 2001) as given by Hampel et al., (1986). The details of implementation are explained as:

$$\text{pru}_{\text{new}} = \frac{1}{2} \left\{ \text{tanh} \left( 0.01 \frac{\text{pru}_{\text{old}} - \text{m}_{\text{pru}}}{\sigma_{\text{pru}}} \right) + 1 \right\}$$
Where, \( \mu_{\text{pro}} \) and \( \sigma_{\text{pro}} \) represents the normalized and original physicochemical property of i-th of amino acid respectively, and \( \mu_{\text{ori}} \) and \( \sigma_{\text{ori}} \) represents the mean and standard deviation of each physicochemical property.

4.5.5 Extraction of feature

4.5.5.1 Feature set 1

Feature set 1 was extracted through following steps:
Step 1: For a particular short peptide sequence (SPS) representing as possible candidate for epitope (EP) or non-epitope (NEP) classes, normalized values of 6 physicochemical properties of each of the residues were listed row-wise in which each of the column represented a particular physicochemical property. This two dimensional array was designated as \( \text{MAT1} = \{P_{ij}\}_{i=1}^{N_6} \) where \( N \) is the length of the SPS.
Step 2: The column-wise minimum, maximum and mean values for each physicochemical property of \( \text{MAT1} \) was calculated to get a one dimensional array (row-vector) \( \text{ARR1} \) of length \( 6 \times 3 = 18 \).
Step 3: The percentage content of each type of amino acids present within the SPS (say \( K \) types of amino acids out of 20 possible) of length \( N \) was calculated as: \( (\text{frequency of amino acid type})/N \) to get a 1 dimensional array \( \text{ARR2} \) of length \( K \). Subsequently, minimum, maximum and mean values of \( \text{ARR2} \) were calculated to keep in another one dimensional array (row-vector), \( \text{ARR3} \) of length 3.
Step 4: The percentage content of each type of dipeptides present within the SPS (say \( L \) types of dipeptides out of total 400 possible) was calculated as: \( (\text{frequency of dipeptide type})/400 \) to get a one dimensional array \( \text{ARR4} \) of length \( L \). Subsequently, minimum, maximum and mean values of \( \text{ARR4} \) were calculated to keep in another one dimensional array (row-vector), \( \text{ARR5} \) of length 3.
Step 5: Finally row-vectors \( \text{ARR1}, \text{ARR3} \) and \( \text{ARR5} \) were concatenated to get final Feature set 1.

4.5.5.2 Feature set 2

To explore effect of dominance of one sequence property over another unrelated sequence property in discriminating EP from NEP, statistics of unrelated properties (as used to form
feature set 1) was measured to extract feature set 2. Feature set 2 was extracted through following steps:

Step 1: For a particular SPS, normalized values of 6 physicochemical properties of each of the residues were concatenated within a row-vector ARR6 of size $N \times 6$ where $N$ is the length of the SPS.

Step 2: The percentage content of each type of amino acids present within the SPS (say $K$ types of amino acids out of 20 possible) of length $N$ was calculated as: (frequency of amino acid type)/$N$ to get a 1 dimensional array ARR7 of length $K$.

Step 3: The percentage content of each type of dipeptides present within the SPS (say $L$ types of dipeptides out of total 400 possible) was calculated as: (frequency of dipeptide type)/400 to get a one dimensional array ARR8 of length $L$.

Step 4: Row-vectors ARR6, ARR7 and ARR8 were concatenated to get a one dimensional array ARR9 of length $(N \times 6) + K + L$.

Step 5: Finally, we calculated length, maximum, minimum, standard deviation, mean and quartile of ARR9 to get Feature set 2 for each SPS.

4.5.6 Application of Elman back propagation

Elman back propagation neural network of Matlab (nntool) was used to evaluate the model performance for prediction of B-cell linear epitopes using Feature sets 1 and 2. For this purpose, training and test set were created for 10 fold cross validation. All the training and test data were created after random reshuffling. For Feature sets 1 and 2, numbers of input nodes for Elman neural network were 24 and 6 respectively. Also, for Feature sets 1 and 2, number of hidden layers were kept as 2 and in each of them the number of nodes were used as 7 and 3 respectively following the formula provided by Shibata and Ikeda (Shibata and Ikeda, 2009).

The details for calculation of nodes in hidden layers are as:

$$N_h = \sqrt{N_i N_o}$$

Where $N_h, N_i$ and $N_o$ representing the number of nodes in hidden layers, input layer and output layer respectively.

4.5.7 Benchmarking with Singh et al., (2013) work

Performance parameters of this work, such as, accuracy, sensitivity, specificity and MCC resulted through section 2.1 to 2.6 were compared against the work done by Singh et al.,(Singh et al., 2013) using common LBtope confirmed positive and negative dataset.