Chapter 4 Techniques for designing viral peptide vaccines

4.1 Introduction
The new paradigm in vaccine design is emerging, following essential discoveries in immunology and development of new major histocompatibility complex (MHC) Class-I binding peptides prediction tools. MHC molecules are cell surface glycoproteins, which take active part in host immune reactions. The involvement of MHC class-I in response to almost all antigens and the variable length of interacting peptides make the study of MHC Class I molecules very interesting. MHC molecules have been well characterized in terms of their role in immune reactions. They bind to some of the peptide fragments generated after proteolytic cleavage of antigen. This binding acts like red flags for antigen specific and to generate immune response against the parent antigen. So a small fragment of antigen can induce immune response against whole antigen. This theme is implemented in designing subunit and synthetic peptide vaccines. In this study we now report on the binding ability of antigenic peptides to MHC class-I, which integrated prediction of peptide MHC class binding; proteasomal C terminal cleavage and TAP transport efficiency of protein.

4.2 Sequence analysis
Sequence based searching is important key skill for biologists. A little exploration of the biological databases at the beginning of a project often saves a lot of time in the lab. Identifying homologous sequences provides bases for phylogenetic analysis and sequence pattern recognition [1]. Sequence based searching can be done online through web applications, so it required special computing skill, but to judge the quality of search results one needs to understand how underlined sequence alignments methods works and go beyond simple sequence alignment to other type of analysis [2-5].

The Multiple Sequence Alignment (MSA) is a set of sequences viewed as an evolutionary history of the sequences. It provides information as to most alike region in the set. In protein such regions represent conserved functional or structural domains. Multiple alignments of protein sequences are important tools in studying sequences. The basic information they provide is identification of conserved sequence regions. This is very useful in designing experiments to test and modify the function
of specific proteins, in predicting the function and structure of proteins, and in identifying new members of protein families [6].

Key Notes –

ℜ Identify all matching words, comprising identical residues, between the sequences.
ℜ Identify high-scoring segments, above a given threshold, occurring within an allowed distance from another high-scoring segment on the same diagonal.
ℜ Extended these segments to include similar residues, using an appropriate substitution matrix.
ℜ Use limited dynamic programming to join the highest scoring segments.
ℜ Multiple sequence alignment of set sequences can provide information as to the most alike region in the set.
ℜ In protein sequences, such region represent conserved functional or structural domains.
ℜ The alignment of a pair of protein sequences reveals whether or not there is an evolutionary relationship among multiple sequences.

A. Multiple sequence alignment of potyviridae family plant viruses-

Potyvirus
The genus Potyvirus is the largest plant virus group comprising more than 30% of all known plant viruses and is economically very important. The genus Potyvirus contains more than 200 members or possible members and belongs to the largest plant virus family, Potyviridae. Potyviruses infect a broad range of monocot and dicot plants and can be responsible for severe damage to crops. Potyviruses have flexuous filamentous particles that contain an approximately 10 kb positive-sense single-stranded RNA that is covalently linked to a virus genome-linked protein (VPg) at the 5’ end and polyadenylated at the 3’ end [7]. The RNA has a single open reading frame that is translated into a large polyprotein, which is proteolytically cleaved into mature proteins by three virus-encoded proteinases. Systemic infection of plants occurs when a virus establishes genome amplification, then movement cell-to-cell and over long distances via the plasmodesmata and phloem cells. The factors that influence virus
host range, symptomatology and pathogenicity have been studied using various virus mutants as well as recombinants constructed from closely related viruses [8, 9]. In many cases these factors involve one or more proteins involved in virus replication and/or transport, including coat protein (CP), movement protein and/or other proteins conferring the function necessary for virus movement [10].

**Infection cycle**

Plant viruses must accomplish three main steps to complete their infection cycle: (i) replication inside the cell, (ii) cell-to-cell movement through plasmodesmata, and (iii) long-distance movement through the vascular tissue. It is the exception rather than the rule that viruses can successfully infect plants, and it is thought that, in most cases, host range limitations are the result of restrictions to virus movement rather than to the inability to replicate within cells [11]. Viral proteins involved in cell-to-cell movement have been well characterized, specially the movement protein. Usually, the coat protein is essential for efficient movement of plant viruses outside the inoculated leaves, although this is not a general rule. Moreover, there is genetic evidence indicating that MPs may perform specific long-distance movement functions, Proteins other than the CPs and the MPs, such as the potyviral HC protein (HC-Pro), the genome-linked protein (VPg) [12-14].

**Movement**

The major contribution of movement defects to host range restrictions suggests that specific host factors could play key roles in virus movement. Both cell-to-cell and long-distance movements have been shown to depend on interactions between virus proteins and species-specific plant factors [15, 16]. However, in spite of the extensive divergence among different virus groups of the movement mechanisms derived from these interactions, they appear to be able to facilitate the transport of a wide range of heterologous viruses. Thus, many viruses belonging to different taxonomic groups have been shown to be able to complement one another’s movement functions in non host plants [17]. Moreover, several studies have shown that different viral proteins, expressed from transgenes, heterologous sequences cloned in defective genomes, or cotransfected plasmids, can functionally replace nonhomologous proteins from other viruses, sometimes resulting in an extension of their virus host ranges [18].
Materials and Methods
Phylogenetics assumed to be an estimate of a phylogeny; branch lengths are proportional to the amount of inferred evolutionary change. Thus cladograms show common ancestry, but do not indicate the amount of evolutionary ‘time’ separating taxa [19, 20]. Here we take sixteen species from potyviridae family of plant virus and study the evolutionary analysis of a family of related protein sequences is determined by phylogenic analysis. Placing the sequences as outer branches on a phylogenetic tree depicted the evolutionary relationships among the potyviridae coat protein sequences. The evolutionary mechanisms giving rise to changes in coat protein sequences and the challenges faced when aligning protein sequences are discussed [21]. Blocks were constructed from a set of align coat protein sequence pairs. Motif analyses of thirteen sequences are representing four blocks, which are conserved region in the multiple sequence alignment.

Results and Interpretations
Here we take sixteen species from potyviridae family of plant virus and evolutionary analysis of a family of related protein sequences is determined by phylogenic analysis. Placing the sequences as outer branches on a phylogenetic tree depicted the evolutionary relationships among the potyviridae coat protein sequences. The evolutionary mechanisms giving rise to changes in coat protein sequences and the challenges faced when aligning protein sequences are discussed. Input coat protein sequences were studied as-
16 Protein Sequence name are as-

>gi130497|sp|P20234.1|POLG_OMV Genome polyprotein [Contains: Viral genome-linked protein (VPg); Nuclear inclusion protein A (NI-A) (NIA) (49 kDa proteinase) (49 kDa-Pro); Nuclear inclusion protein B (NI-B) (NIB) (RNA-directed RNA polymerase); Coat protein (CP)]

>gi221058|dbj|BAA01741.1| polyprotein precursor [Barley mild mosaic virus]

>gi221426|dbj|BAA01892.1| polyprotein precursor [Leek yellow stripe virus]

>gi497916|gb|AAB50167.1| polyprotein

>gi632065|pir|S45323 genome polyprotein 1 - wheat yellow mosaic virus (isolate South Trance) (fragment)

>gi1181180|dbj|BAA01742.1| polyprotein precursor [Barley mild mosaic virus]

>gi1304228|dbj|BAA01892.1| polyprotein [Sweet potato feathery mottle virus]

>gi2554632|dbj|BAA22880.1| polyprotein [Bean yellow mosaic virus]

>gi2952295|gb|AAC05494.1| polyprotein [Dasheen mosaic virus]

>gi4033458|sp|Q65729.1|POLG_BSTVG Genome polyprotein [Contains: Nuclear inclusion protein A (NI-A) (NIA) (49 kDa proteinase) (49 kDa-Pro); Nuclear inclusion protein B (NI-B) (NIB) (RNA-directed RNA polymerase); Coat protein (CP)]

>gi11559225|dbj|BAA18744.1| 270K polyprotein [Barley yellow mosaic virus]

>gi19744020|gb|CAA76842.3| polyprotein [Sugarcane streak mosaic virus]

>gi2819134|gb|AAO33413.1|AF469171_1 polyprotein precursor [Calla lily latent virus]

>gi68235818|gb|AAY88245.1| polyprotein precursor [Bean common mosaic virus]

>gi76555913|gb|CAI23783.1| polyprotein [Shallot yellow stripe virus]

>gi14152858|sp|P18478.3|POLG_WMV2U Genome polyprotein [Contains: Nuclear inclusion protein A (NI-A) (NIA) (49 kDa proteinase)
Fig. 4.1- Multiple Sequence Alignments for polyprotein from 16 different species of potyviridae family
Interpretation of ClustalW

An alignment will display by default the following symbols denoting the degree of conservation observed in each column.

1. ‘*’ Means that, the residues in that column are identical in all sequences in the alignment.
2. ‘:’ Means that, conserved substitutions have been observed, according to the COLOUR table.
3. ‘.’ Means that, semi-conserved substitutions are observed.

Application of Multiple sequence alignment-

1. Identification for functionally important sites.
2. Demonstration of homology between sequences.
3. Molecular phylogeny to retrace evolutionary relationships between sequences.
4. Search for weak but significant similarities in sequence database.
5. Function prediction – similar shapes usually implies similar function.

B. Multiple sequence alignment of Mungbean yellow mosaic India virus - Geminiviruses family

- Indian Mungbean yellow mosaic virus
- Mungbean yellow mosaic India virus - [Akola]
- Mungbean yellow mosaic India virus - [Cowpea Pakistan]
- Mungbean yellow mosaic India virus - [India:Varanasi:Dolichos:2005]
- Mungbean yellow mosaic India virus - [Mungbean Pakistan]
- Mungbean yellow mosaic India virus - [Nepal]
- Mungbean yellow mosaic India virus-[Bangladesh]
- Mungbean yellow mosaic India virus-[Bogra]
- Mungbean yellow mosaic India virus-[Cowpea]
- Mungbean yellow mosaic India virus-[Mungbean]
- Mungbean yellow mosaic India virus-[Soybean]
- Mungbean yellow mosaic India virus-[SoybeanTN]
Fig. 4.2a - Multiple Sequence Alignments of different strains for coat protein from *Mungbean yellow mosaic India virus*
**Fig. 4.2c- Multiple Sequence Alignments of different strains for coat protein from Mungbean yellow mosaic India virus**

**Interpretation**

- Phylogenetics analysis of *geminiviruses* assumed to be an estimate of a phylogeny; branch lengths are proportional to the amount of inferred evolutionary change.
- Transport of the viral genome into the nucleus is an obligatory step in the replication cycle of plant geminiviruses.
- In these virus types, the multifunctional coat protein (CP) of *Indian mungbean yellow mosaic virus* (MYMV) is thought to be involved in this process.
- Given a set of gene sequences, it should be possible to reconstruct the evolutionary relationships among genes and among organisms.
- The evolutionary mechanisms giving rise to changes in coat protein sequences and the challenges faced when aligning protein sequences are discussed.
C. Phylogenetic analysis of thirty species from Potyviridae Taxa:

Phylogenetics assumed to be an estimate of a phylogeny; branch lengths are proportional to the amount of inferred evolutionary change. The phylogenetics study tree-building methods are varying in many properties. Some methods choose the tree from among all of those possible that either maximises or minimises some optimality criterion, while other methods merely follow an algorithm (a pre-defined sequence of operations) to produce a single tree (constructive methods). Some of the methods analyse the data matrix directly, while others require that the data have been converted to distances between taxa before the analysis i.e. a matrix of all possible pair-wise distances between the taxa is calculated. Furthermore, some of the methods that use an optimality criterion evaluate the optimality criterion over all taxa, while some methods evaluate the criterion on subsets of the taxa usually four taxa, called quartets. Some methods have been developed specifically for sequence data, while others may be applied to any type of data. The more commonly used methods and their characteristics are discussed below. For those methods that have an optimality criterion, it is important to recognise that they are based on a double-level optimisation, where first the optimality criterion is optimised for a given tree topology, and then the tree that optimises these optimal values is selected i.e. first you assess how good each tree is, and then you find the tree that is the best. Methods thus need to be known that accomplish the optimisation for a specified tree, and then methods need to be known for searching among the trees for the optimal one. The first class of methods will be unique for each tree-building procedure, but the second class of methods is more general and can be used for all tree-building methods [22].

Several strategies for finding the optimal tree or trees and exact methods are those that guarantee to find the optimal solution, while heuristic methods are computationally efficient strategies that should produce a solution that is at least close to the optimal one even if it doesn't find the optimum. The most efficient of the exact methods involve the use of the branch-and-bound strategy. The need for heuristic methods is simply one of practicality - the number of possible trees to be tested by the exact methods increases exponentially with increasing numbers of taxa. Heuristic search strategies first try to find a tree that is a good estimate of the final solution, usually by sequentially adding the taxa to the growing tree, adding each taxon in the optimal place on the tree. Most heuristic strategies then search through those trees that are
similar to this initial tree in order to find a better solution, usually using branch swapping. The heuristic searches do not necessarily find the optimal solution, nor do they necessarily get particularly close to it. It is therefore important to accurately describe any heuristic procedures used in an analysis, so that the likely success of the method can be assessed [23].

**Neighbor-joining analysis**

Neighbor-joining analysis is based on the minimum-evolution criterion for phylogenetic trees, i.e. the topology that gives the least total branch length is preferred at each step of the algorithm. However, neighbor-joining is sub-optimal in this sense, it has been extensively tested and usually finds a tree that is quite close to the optimal tree.

Neighbor-joining is a method that is related to the cluster method but does not require the data to be ultrametric. In other words it does not require that all lineages have diverged by equal amounts. The method is especially suited for datasets comprising lineages with largely varying rates of evolution. It can be used in combination with methods that allow correction for superimposed substitutions [24].

The neighbor-joining method is a special case of the star decomposition method. In contrast to cluster analysis neighbor-joining keeps track of nodes on a tree rather than taxa or clusters of taxa. The raw data are provided as a distance matrix and the initial tree is a star tree. Then a modified distance matrix is constructed in which the separation between each pair of nodes is adjusted on the basis of their average divergence from all other nodes. The tree is constructed by linking the least-distant pair of nodes in this modified matrix. When two nodes are linked, their common ancestral node is added to the tree and the terminal nodes with their respective branches are removed from the tree. This pruning process converts the newly added common ancestor into a terminal node on a tree of reduced size. At each stage in the process two terminal nodes are replaced by one new node. The process is complete when two nodes remain, separated by a single branch [25, 26].
Phylogenetic analysis-

- Phylogram is a branching diagram (tree) assumed to be an estimate of a phylogeny; branch lengths are proportional to the amount of inferred evolutionary change.
- A Cladogram is a branching diagram (tree) assumed to be an estimate of a phylogeny where the branches are of equal length, thus cladogram show common ancestry, but do not indicate the amount of evolutionary ‘time’ separating taxa.
- Tree distances can be shown. Just click on the diagram to get a menu of options. The ".dnd" file is a file that describes the phylogenetic tree.

Key Notes

- The evolutionary analysis of a family of related protein sequences is determined by phylogenetic analysis.
- Placing the sequences as outer branches on a tree depicts the evolutionary relationships among the sequences.
- The evolutionary mechanisms giving rise to changes in protein sequences and the challenges faced when aligning protein sequences are discussed.
- Phylogenetics assumed to be an estimate of a phylogeny; branch lengths are proportional to the amount of inferred evolutionary change in coat proteins.
- The evolutionary relationships among the coat proteins sequences were depicted by placing the sequences as outer branches on a tree; here we take thirty species from of plants and evolutionary analysis of a family of related maturase protein sequences is determined by phylogenic analysis.
- The evolutionary mechanisms giving rise to changes in coat proteins sequences and the challenges faced when aligning protein sequences are discussed. Blocks were constructed from a set of align coat proteins sequence pairs. Motif analyses of thirty sequences are representing five blocks- a conserved region in the multiple sequence alignment.
Table 4.1- information of thirty species from Potyviridae Taxa

<table>
<thead>
<tr>
<th>Serial No.</th>
<th>Name</th>
<th>GI Number</th>
<th>Accession Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Oat mosaic virus</td>
<td>gi</td>
<td>25014013</td>
</tr>
<tr>
<td>2</td>
<td>Wheat spindle streak mosaic virus</td>
<td>gi</td>
<td>74095603</td>
</tr>
<tr>
<td>3</td>
<td>Wheat yellow mosaic virus</td>
<td>gi</td>
<td>25013965</td>
</tr>
<tr>
<td>4</td>
<td>Cassava brown streak virus</td>
<td>gi</td>
<td>112554420</td>
</tr>
<tr>
<td>5</td>
<td>Sweet potato mild mosaic virus</td>
<td>gi</td>
<td>25013877</td>
</tr>
<tr>
<td>6</td>
<td>Indian cardamom mosaic virus</td>
<td>gi</td>
<td>13810302</td>
</tr>
<tr>
<td>7</td>
<td>Alstroemeria mosaic virus</td>
<td>gi</td>
<td>50058108</td>
</tr>
<tr>
<td>8</td>
<td>Amazon lily mosaic virus</td>
<td>gi</td>
<td>50058110</td>
</tr>
<tr>
<td>9</td>
<td>Banana bract mosaic virus</td>
<td>gi</td>
<td>152926136</td>
</tr>
<tr>
<td>10</td>
<td>Blackeye cowpea mosaic virus</td>
<td>gi</td>
<td>82571432</td>
</tr>
<tr>
<td>11</td>
<td>Peanut stripe virus</td>
<td>gi</td>
<td>87243812</td>
</tr>
<tr>
<td>12</td>
<td>Dendrobium mosaic virus</td>
<td>gi</td>
<td>773364</td>
</tr>
<tr>
<td>13</td>
<td>Beet mosaic virus</td>
<td>gi</td>
<td>40254038</td>
</tr>
<tr>
<td>14</td>
<td>Cymbidium potyvirus</td>
<td>gi</td>
<td>57162045</td>
</tr>
<tr>
<td>15</td>
<td>Carnation vein mottle virus</td>
<td>gi</td>
<td>28975731</td>
</tr>
<tr>
<td>16</td>
<td>Ageratum houstonianum potyvirus</td>
<td>gi</td>
<td>56797907</td>
</tr>
<tr>
<td>17</td>
<td>Vanilla distortion mosaic virus</td>
<td>gi</td>
<td>95007591</td>
</tr>
<tr>
<td>18</td>
<td>Triticum mosaic virus</td>
<td>gi</td>
<td>133900018</td>
</tr>
<tr>
<td>19</td>
<td>Sugarcane streak mosaic virus</td>
<td>gi</td>
<td>170015557</td>
</tr>
<tr>
<td>20</td>
<td>Chinese yam necrotic mosaic virus</td>
<td>gi</td>
<td>145586909</td>
</tr>
<tr>
<td></td>
<td>Virus Name</td>
<td>Accession</td>
<td>Protein Accession</td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------</td>
<td>------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>21</td>
<td>Cucumber vein yellowing virus</td>
<td>gi</td>
<td>149227982</td>
</tr>
<tr>
<td>22</td>
<td>Oat necrotic mottle virus</td>
<td>gi</td>
<td>38304210</td>
</tr>
<tr>
<td>23</td>
<td>Wheat streak mosaic virus</td>
<td>gi</td>
<td>159141808</td>
</tr>
<tr>
<td>24</td>
<td>Ryegrass mosaic virus</td>
<td>gi</td>
<td>25013868</td>
</tr>
<tr>
<td>25</td>
<td>Hordeum mosaic virus</td>
<td>gi</td>
<td>51101437</td>
</tr>
<tr>
<td>26</td>
<td>Agropyron mosaic virus</td>
<td>gi</td>
<td>50404822</td>
</tr>
<tr>
<td>27</td>
<td>Dianella chlorotic mottle virus</td>
<td>gi</td>
<td>68160331</td>
</tr>
<tr>
<td>28</td>
<td>Florida passionflower potyvirus LAJ-2006</td>
<td>gi</td>
<td>12820164</td>
</tr>
<tr>
<td>29</td>
<td>Hardenbergia mosaic virus</td>
<td>gi</td>
<td>148841017</td>
</tr>
<tr>
<td>30</td>
<td>Nerine potyvirus IVT80054</td>
<td>gi</td>
<td>125505647</td>
</tr>
</tbody>
</table>
4.3- Rooted tree built for Neighbor-joining analysis of potyviridae family polyprotein showing the evolutionary difference
4.4- Unrooted tree builded for Neighbor-joining analysis of potyviridae family polyprotein showing the evolutionary difference
4.5- Rooted dendrogram builded for Neighbor-joining analysis of potyviridae family polyprotein showing the evolutionary difference
4.6- Unrooted dendrogram builded for Neighbor-joining analysis of potyviridae family polyprotein showing the evolutionary difference
D. Phylogenetic analysis of thirty species from *Tombusviridae* taxonomy:

Table 4.2- information of thirty species from *Tombusviridae* taxonomy

<table>
<thead>
<tr>
<th>Serial No.</th>
<th>Name</th>
<th>GI Number</th>
<th>Accession Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cucumber leaf spot virus</td>
<td>gi</td>
<td>898888604</td>
</tr>
<tr>
<td>2</td>
<td>Turnip crinkle virus</td>
<td>gi</td>
<td>116807</td>
</tr>
<tr>
<td>3</td>
<td>Johnsongrass chlorotic stripe mosaic virus</td>
<td>gi</td>
<td>39163651</td>
</tr>
<tr>
<td>4</td>
<td>Maize white line mosaic virus</td>
<td>gi</td>
<td>87243806</td>
</tr>
<tr>
<td>5</td>
<td>Sesame necrotic mosaic virus</td>
<td>gi</td>
<td>20177494</td>
</tr>
<tr>
<td>6</td>
<td>Oat chlorotic stunt virus</td>
<td>gi</td>
<td>9627424</td>
</tr>
<tr>
<td>7</td>
<td>Cardamine chlorotic fleck virus</td>
<td>gi</td>
<td>76780417</td>
</tr>
<tr>
<td>8</td>
<td>Angelonia flower mottle virus</td>
<td>gi</td>
<td>116793</td>
</tr>
<tr>
<td>9</td>
<td>Red clover necrotic mosaic virus Hibiscus chlorotic ringspot virus</td>
<td>gi</td>
<td>171903995</td>
</tr>
<tr>
<td>10</td>
<td>Maize chlorotic mottle virus</td>
<td>gi</td>
<td>169410535</td>
</tr>
<tr>
<td>11</td>
<td>Melon necrotic spot virus</td>
<td>gi</td>
<td>145588174</td>
</tr>
<tr>
<td>12</td>
<td>Pea stem necrosis virus</td>
<td>gi</td>
<td>32469482</td>
</tr>
<tr>
<td>13</td>
<td>Pelargonium flower break virus</td>
<td>gi</td>
<td>90296772</td>
</tr>
<tr>
<td>14</td>
<td>Angelonia flower break virus</td>
<td>gi</td>
<td>85718598</td>
</tr>
<tr>
<td>15</td>
<td>Sikte waterborne virus</td>
<td>gi</td>
<td>46560514</td>
</tr>
<tr>
<td>16</td>
<td>COAT_TBSVB Coat protein</td>
<td>gi</td>
<td>116805</td>
</tr>
<tr>
<td>17</td>
<td>COAT_TBSVC Coat protein (P41 capsid protein)</td>
<td>gi</td>
<td>116806</td>
</tr>
<tr>
<td>18</td>
<td>Petunia asteroid mosaic virus</td>
<td>gi</td>
<td>46560498</td>
</tr>
<tr>
<td>19</td>
<td>Pelargonium leaf curl virus</td>
<td>gi</td>
<td>299097</td>
</tr>
<tr>
<td></td>
<td>Virus Name</td>
<td>Accession</td>
<td>Species</td>
</tr>
<tr>
<td>---</td>
<td>------------------------------------------------</td>
<td>-----------</td>
<td>---------------</td>
</tr>
<tr>
<td>20</td>
<td>Pear latent virus</td>
<td>gi</td>
<td>30018249</td>
</tr>
<tr>
<td>21</td>
<td>Neckar river virus</td>
<td>gi</td>
<td>46560510</td>
</tr>
<tr>
<td>22</td>
<td>Moroccan pepper virus</td>
<td>gi</td>
<td>23344113</td>
</tr>
<tr>
<td>23</td>
<td>Grapevine Algerian latent virus</td>
<td>gi</td>
<td>53692199</td>
</tr>
<tr>
<td>24</td>
<td>COAT_CRV Coat protein</td>
<td>gi</td>
<td>116738</td>
</tr>
<tr>
<td>25</td>
<td>COAT_CNV Coat protein</td>
<td>gi</td>
<td>116737</td>
</tr>
<tr>
<td>26</td>
<td>Cucumber Bulgarian virus</td>
<td>gi</td>
<td>28274774</td>
</tr>
<tr>
<td>27</td>
<td>Carnation Italian ringspot virus</td>
<td>gi</td>
<td>46560490</td>
</tr>
<tr>
<td>28</td>
<td>Artichoke mottled crinkle virus</td>
<td>gi</td>
<td>58633</td>
</tr>
<tr>
<td>29</td>
<td>COAT_TBSVC Coat protein (P41 capsid protein)</td>
<td>gi</td>
<td>116806</td>
</tr>
<tr>
<td>30</td>
<td>Sikte waterborne virus</td>
<td>gi</td>
<td>46560514</td>
</tr>
</tbody>
</table>
4.7- Rooted tree builded for Neighbor-joining analysis of tombusviridae family coat protein showing the evolutionary difference
4.8- Unrooted tree builded for Neighbor-joining analysis of tombusviridae family coat protein showing the evolutionary difference
4.9- Rooted dendrogram builded for Neighbor-joining analysis of tombusviridae family coat protein showing the evolutionary difference
4.10- Unrooted dendrogram built for Neighbor-joining analysis of tombusviridae family coat protein showing the evolutionary difference

Neighbor-joining method is proposed for reconstructing phylogenetic trees from evolutionary distance of coat proteins from Tombusviridae family. The principle of this method is to find pairs of operational taxonomic units (OTUs [= neighbors]) that minimize the total branch length at each stage of clustering of OTUs starting with a starlike tree. The branch lengths as well as the topology of a parsimonious tree can quickly be obtained by using this method. Using computer simulation, we studied the
efficiency of this method in obtaining the correct unrooted tree in comparison with that of five other tree-making methods: the unweighted pair group method of analysis, Farris's method, Sattath and Tversky's method, Li's method, and Tateno et al.'s modified Farris method. The new, neighbor-joining method and Sattath and Tversky's method are shown to be generally better than the other methods.

**Negative branch lengths**
As the neighbor-joining algorithm seeks to represent the data in the form of an additive tree, it can assign a negative length to the branch. Here the interpretation of branch lengths as an estimated number of substitutions gets into difficulties. When this occurs it is advised to set the branch length to zero and transfer the difference to the adjacent branch length so that the total distance between an adjacent pair of terminal nodes remains unaffected. This does not alter the overall topology of the tree [27].

**Advantages and disadvantages of the neighbor-joining method**
- **Advantages**
  - is fast and thus suited for large datasets and for bootstrap analysis
  - permist lineages with largely different branch lengths
  - permits correction for multiple substitutions
- **Disadvantages**
  - sequence information is reduced
  - gives only one possible tree
  - strongly dependent on the model of evolution used.

**Interpretation of results**
Phylogenetics assumed to be an estimate of a phylogeny; branch lengths are proportional to the amount of inferred evolutionary change in potyviruses. Blocks were constructed from a set of align coat protein sequence pairs. Motif analyses of seventeen sequences are representing four blocks, which shows the distinct amino acids and some region are identical which are conserved from these species which having the phylogenetic resemble from the same family. These patterns include a region with highest 47 matching character followed by a short spacer region until the
sequences start to be different. These patterns are all of the same length and they are aligned, the matching sequence characters will appear in columns.

**Conclusion**

The evolutionary relationships among the potyviridae coat protein sequences were depicted, which shows changes in coat protein sequences and the challenges faced when different pesticide are used for the paste control because of these mutation of same coat protein sequences get different results on plant to protect from the viral attack. This work provides insight into the evolutionary history for the gene families of the potyviridae, linking the expansion of these families to the duplications of the gene cluster regions, and showing that they are composed of subgroups with distinct evolutionary (and possibly functional) differences. Such regions represent conserved functional or structural domains in predicting the function and structure of proteins from ssRNA positive-strand potyviruses, and in identifying new members of protein families.

**Application of Phylogenetic analysis**

- Identification for functionally important sites.
- Demonstration of homology between sequences.
- Molecular phylogeny to retrace evolutionary relationships between sequences.
- Search for weak but significant similarities in sequence database.
- Increases significantly the efficiency of protein secondary structure prediction.
- Function prediction – similar shape usually implies similar function.
- Design of primer for PCR (polymerase chain reaction) identifies of related genes.
4.3 Prediction of antigenic peptides

The protein sequence of plant viruses were analyzed and characterized to study the antigenicity, solvent accessible regions and MHC class peptide binding, which allows potential drug targets to identify active sites against allergic reactions. Prediction of antigenicity program predicts those segments from within protein that are likely to be antigenic by eliciting an antibody response. Antigenic epitopes is determined using the Gomase and Kale (2007), Hopp and Woods antigenicity, Welling antigenicity, Parker antigenicity, Protrusion Index antigenicity, B-EpiPred Server antigenicity, Kolaskar and Tongaonkar antigenicity, Emini surface activity, Karplus-Schulz flexibility prediction methods [28-36]. Predictions are based on plots that reflect the occurrence of amino acid residues in experimentally known segmental epitopes.

A alfalfa mosaic virus coat protein sequence is 221 residues long as-

MSSSQKAGGKAGKPTKRSQNYAALRKAQLPKPPALKVPVVKPTNTILPQTG
CVWQLGTPSLSFNGLGVRFLYSFLKDFAFSLEDLIYRMVFSITPSYAG
TFCLTDVTTEGDRAVAHGNPMQEPHPGAFHANEKFGFELVFTAPTHAGMQ
NQNFKHSYAVALCLDFDAQPEGSKNPSYRFNEVWVERKAFPRAGPLRSLLTV
GLLDEADDLDRH

In these methods, we found the antigenic determinants by finding the area of greatest local hydrophilicity. The Hopp-Woods scale was designed to predict the locations of antigenic determinants in a protein, assuming that the antigenic determinants would be exposed on the surface of the protein and thus would be located in hydrophilic regions. Its values are derived from the transfer-free energies for amino acid side chains between ethanol and water. Welling antigenicity plot gives value as the log of the quotient between percentage in a sample of known antigenic regions and percentage in average proteins.
A method is presented for locating protein antigenic determinants by analyzing amino acid sequences in order to find the point of greatest local hydrophilicity. This is accomplished by assigning each amino acid a numerical value (hydrophilicity value) and then repetitively averaging these values along the peptide chain. The point of highest local average hydrophilicity is invariably located in, or immediately adjacent to, an antigenic determinant. It was found that the prediction success rate depended on averaging group length, with alfalfa mosaic virus peptide averages yielding optimal results. The method was developed using coat protein for which extensive immunochemical analysis has been carried out and subsequently was used to predict antigenic determinants for the coat protein surface antigen, human histocompatibility antigen MHC-I and MHC-II. The coat protein surface antigen sequence was analyzed by physiochemical means and was shown to have antigenic activity by radioimmunoassay.

Fig. 4.12- Welling antigenicity plot gives value as the log of the quotient between percentage in a sample of known antigenic regions and percentage in average proteins.
Prediction of antigenic regions in a coat protein will be helpful for a rational approach to the synthesis of peptides which may elicit antibodies reactive with the intact protein. Earlier methods are based on the assumption that antigenic regions are primarily hydrophilic regions at the surface of the protein molecule. The method presented here is based on the amino acid composition of known antigenic regions in coat protein, which is compared with other proteins. Antigenicity values were derived from the differences between the two data sets. The method was applied to good correlation between the predicted regions and previously determined antigenic regions.

Fig. 4.13- Parker antigenicity plot of coat protein

A new set of hydrophilicity high-performance liquid chromatography (HPLC) parameters is presented. These parameters were derived from the retention times of 20 model synthetic peptides, Ac-Gly-X-X-(Leu)3-(Lys)2-amide, where X was substituted with the 20 amino acids found in coat protein. Since hydrophilicity parameters have been used extensively in algorithms to predict which amino acid residues are antigenic, we have compared the profiles generated by our new set of hydrophilic HPLC parameters on the same Parker scale as nine other sets of parameters. Generally, it was found that the HPLC parameters obtained in this study correlated best with antigenicity. In addition, it was shown that a combination of the three best parameters for predicting antigenicity further improved the predictions.

These predicted surface sites of coat protein or, in other words, the hydrophilic, accessible, or mobile regions were then correlated to the known antigenic sites from immunological studies and accessible sites determined by X-ray crystallographic data for coat protein.
The best single method for predicting linear B-cell epitopes is the hidden Markov model. Combining the hidden Markov model with one of the best propensity scale methods, we obtained the BepiPred method. When tested on the validation data set this method performs significantly better than any of the other methods tested.

Table 4.3- Bepipred predicted epitopes of coat protein from *Alfalfa mosaic virus*

<table>
<thead>
<tr>
<th>No.</th>
<th>Start Position</th>
<th>End Position</th>
<th>Peptide</th>
<th>Peptide Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>21</td>
<td>MSSSQKKAGGKAGKPTKRSQN</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>37</td>
<td>QLPKPPALK</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>49</td>
<td>VKPTNTILP</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>87</td>
<td>87</td>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>103</td>
<td>105</td>
<td>SYA</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>113</td>
<td>133</td>
<td>DVTTEDGRAVAGHNPMQEFPH</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>136</td>
<td>137</td>
<td>FH</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>151</td>
<td>161</td>
<td>PTHAGMQNQNF</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>174</td>
<td>186</td>
<td>DAQPEGSKPNPSYR</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>196</td>
<td>202</td>
<td>AFPRAGP</td>
<td>7</td>
</tr>
</tbody>
</table>
Analysis of data from experimentally determined antigenic sites on coat protein has revealed that the hydrophobic residues Cys, Leu and Val, if they occur on the surface of a protein, are more likely to be a part of antigenic sites. A semi-empirical method which makes use of physicochemical properties of amino acid residues and their frequencies of occurrence in experimentally known segmental epitopes was developed to predict antigenic determinants on coat protein. Application of this method to a large number of proteins has shown that our method can predict antigenic determinants with about 75% accuracy which is better than most of the known methods. This method is based on a single parameter and thus very simple to use.

Fig. 4.17-Comparative surface feature analyses of the coat protein sequences of alfalfa mosaic virus allowed an alignment of the two sequences and an identification
of probable alfalfa mosaic virus neutralization antigenic sites. A synthetic peptide containing the alfalfa mosaic virus specific amino acid sequence of one of these sites induced anti-alfalfa mosaic virus-neutralizing antibodies. It is concluded that a structural homology exists between the two viruses, despite minimal primary sequence conservation.

Fig. 4.18- Karplus-Schulz prediction of chain flexibility in coat Protein

Fig. 4.19- Structure of antigenic peptide having sequence – 86-FGRLIPVPVSE-98 of viral coat protein from Tomato aspermy virus
4.4 Prediction of protein secondary structure

The important concepts in secondary structure prediction are: residue conformational propensities, sequence edge effects, moments of hydrophobicity, position of insertions and Deletions in aligned homologous sequence, moments of conservation, auto-correlation, residue ratios, secondary structure feedback effects, and filtering [37-39].
The Robson and Garnier method predicted the secondary structure of coat protein. Each residue is assigned values for alpha helix, beta sheet, turns and coils using a window of 7 residues. Using these information parameters, the likelihood of a given residue assuming each of the four possible conformations alpha, beta, reverse turn, or coils calculated, and the conformation with the largest likelihood is assigned to the residue.

Fig. 4.22- secondary structure of alfalfa mosaic virus coat protein, showing helix, sheet and turns

Fig. 4.23- Secondary structure plot of coat protein

Fig. 4.24- SOPMA: Significant improvement in protein secondary structure prediction by consensus prediction
Recently a new method called the self-optimized prediction method (SOPM) was described to improve the success rate in the prediction of the secondary structure of proteins. We report improvements brought about by predicting all the sequences of a set of aligned proteins belonging to the same family. This improved SOPM method (SOPMA) correctly predicts 69.5% of amino acids for a three-state description of the secondary structure (alpha-helix, beta-sheet and coil) in a whole database containing 126 chains of non-homologous (less than 25% identity) proteins. Joint prediction with SOPMA and a neural networks method (PHD) correctly predicts 82.2% of residues for 74% of co-predicted amino acids.

4.5 Finding the location in solvent accessible regions
Finding the location in solvent accessible regions in protein determines the hydrophobic and hydrophilic scales and is utilized for prediction. This is useful in predicting membrane-spanning domains, potential antigenic sites and regions that are exposed on the protein surface. The MHC peptide binding is predicted using neural networks trained on C terminals of known epitopes. Solvent accessible scales for delineating hydrophobic and hydrophilic characteristics of amino acids and scales are developed for predicting potential antigenic sites of globular proteins, which are rich in charged and polar residues [40-61].

4.5.1 Sweet hydrophobicity

![Sweet hydrophobicity Plot](image)

Fig. 4.25- Sweet hydrophobicity Plot
4.5.2 Kyte & Doolittle hydrophobicity

Fig. 4.26- Kyte & Doolittle hydrophobicity plot

4.5.3 Abraham & Leo hydrophobicity

Fig. 4.27- Abraham & Leo hydrophobicity plot

4.5.4 Bull & Breese hydrophobicity

Fig. 4.28- Bull & Breese hydrophobicity plot
4.5.5 Guy hydrophobicity

Fig. 2.29- Guy hydrophobicity plot

4.5.6 Miyazawa hydrophobicity

Fig. 4.30-Miyazawa hydrophobicity plot

4.5.7 Roseman hydrophobicity

Fig. 4.31-Roseman hydrophobicity plot
4.5.8 Cowan HPLC pH7.5 hydrophobicity

Fig. 4.32-Cowan HPLC pH7.5 hydrophobicity plot

4.5.9 Rose hydrophobicity

Fig. 4.33-Rose hydrophobicity plot

4.5.10 Eisenberg hydrophobicity

Fig. 4.34-Eisenberg hydrophobicity plot
4.5.11 Manavalan hydrophobicity

Fig. 4.35-Manavalan hydrophobicity plot

4.5.12 Black hydrophobicity

Fig. 4.36-Black hydrophobicity plot

4.5.13 Fauchere hydrophobicity

Fig. 4.37-Fauchere hydrophobicity plot
4.5.14 Janin hydrophobicity

![Fig. 4.38-Janin hydrophobicity plot]

4.5.15 Rao & Argos hydrophobicity

![Fig. 4.39- Rao & Argos hydrophobicity plot]

4.5.16 Wolfenden hydrophobicity

![Fig. 4.40- Wolfenden hydrophobicity plot]
4.5.17 Wilson HPLC hydrophobicity

Fig. 4.41- Wilson HPLC hydrophobicity

4.5.18 von Heijne Hydrophilicity

Fig. 4.42- von Heijne Hydrophilicity of coat protein

4.5.19 Chothia hydrophobicity scales

Fig. 4.43- Chothia hydrophobicity scales plot
4.6 MHC class binding peptides

4.6.1 SVM Based Prediction

The new paradigm in vaccine design is emerging, following essential discoveries in immunology and development of new MHC Class-I binding peptides prediction tools. MHC molecules are cell surface glycoproteins, which take active part in host immune reactions [62, 63]. The involvement of MHC class-I in response to almost all antigens and the variable length of interacting peptides make the study of MHC Class I molecules very interesting. MHC molecules have been well characterized in terms of
their role in immune reactions. They bind to some of the peptide fragments generated after proteolytic cleavage of antigen. This binding acts like red flags for antigen specific and to generate immune response against the parent antigen. So a small fragment of antigen can induce immune response against whole antigen. This theme is implemented in designing subunit and synthetic peptide vaccines [64].

In analysis, predicted MHC/peptide binding is a log-transformed value related to the IC50 values in nM units. MHC2Pred predicts peptide binders to MHCI and MHCII molecules from protein sequences or sequence alignments using Position Specific Scoring Matrices (PSSMs). The Support Vector Machine (SVM) based method is for prediction of promiscuous MHC class II binding peptides. The average accuracy of SVM based method for 42 alleles is ~80%. This method is also useful in cellular immunology, Vaccine design, immunodiagnostics, immunotherapeutics and molecular understanding of autoimmune susceptibility. For development of MHC binder, an elegant machine learning technique SVM is used. SVM is trained on the binary input of single amino acid sequence. In addition, we predict those MHC ligands from whose C-terminal end is likely to be the result of proteosomal cleavage. The threshold is used to discriminate the MHC binders from non-binders. The user can vary the threshold score between -1.5 to 1.5. The peptides achieving score more then the cutoff score are predicted as binders otherwise they are predicted as non-binders. If the user did not select any cutoff score then the default threshold of prediction methods will be used. The default threshold is that at which the sensitivity and specificity of prediction methods are nearly same [65].

**Prediction Information-:** The helper T cell epitopes are subset of MHC class II ligands and play a decisive role in initiation and maintenance of immune response. Experimental identification of such ligands is arduous, time-consuming, and economically not feasible. Therefore, development of reliable computational methods for their prediction may reduce cost and number of wet lab experiments to identify these peptides. The prediction of MHC class II binding peptides is difficult as compared to MHC class I binding peptides due to their variable size. MHC class II binding peptides are 10-40 amino acids long with a binding core of 9 amino acids containing primary anchor residues. Therefore, in case of MHC class II binders prediction, an additional method for finding binding core of 9 amino acids from ligands of variable length is required. In the literature, numerous reasonably validated mathematical models for the prediction of core and binding properties of these binding peptides are available. The methods for core prediction are based on genetic programming,
discriminant analysis and matrix optimization techniques. The methods for the prediction of binders are based on motif, quantitative matrices and artificial neural networks. Most of these methods are available for HLA-DRB1*0401 allele, associated with autoimmune disease rheumatoid arthritis. These methods are not able to predict the peptides binding to many MHC alleles or promiscuous MHC binders. In this study, an attempt has been made to develop a highly accurate prediction method for large number of MHC class II alleles. The matrix optimization technique has been used to detect the binding core of peptides. Subsequently support vector machine has been used for discrimination between binding and non-binding peptides. The overall accuracy of method is $>78\%$, which is better than all the already existing methods in literature [66].

**Algorithm for development of prediction method**

The binders and non-binders for all alleles have been obtained from MHCBN and JenPep database. All the peptides having IC50 value less than 500nm has been considered as binders and peptides with IC50 value greater than 500nm are considered as non-binders. Peptides containing less than 9 amino acids have been deleted from the dataset. The binding core of 9 amino acids has been obtained from the binders of variable length without considering MHC binding motifs using Matrix Optimization Techniques (MOT) package. For development of MHC binder prediction method, an elegant machine learning technique SVM has been used. SVM has been trained on the binary input of single amino acid sequence. Each amino acid of 9mer peptide was represented by a 20-dimensional vector. Each peptide of nine amino acids has been represented through a vector of 180 dimensions. The binders have been represented by the +1 and non-binders by -1. A suitable type of kernel for classifying the data has been chosen by conducting experiments with every kernel type i.e. RBF, Polynomial, linear and Sigmoid. The kernel features and regulatory parameter C were optimized by systematic variation in the parameters and evaluations of prediction performance. The overall architecture of SVM based methods is shown in below [67, 68].

**Cross-validation and Performance measures**

The main goal of machine learning approach is to obtain good classification performance on unseen data. Therefore, performance of methods for all alleles has been evaluated using 5-fold cross validation. In 5-fold cross-validation the dataset is randomly divided into five equal sized subsets. The method is trained 5 times using 1 distinct set for testing and remaining 4 sets for training. The final performance of the method is obtained by averaging. The performance of method has been measured.
through threshold dependent parameters such as sensitivity, specificity, NPV, PPV and accuracy.

**Alfalfa mosaic virus MHC Binding peptides**

These MHC binding peptides are sufficient for eliciting the desired immune response. Predicted MHC binding regions in an antigen sequence are directly associated with immune reactions. In analysis we found the MHCI and MHCII binding regions (table -2, 3, and 4). NetMHC 2.1 server predicts binding of peptides to a number of different alleles using artificial neural networks (ANNs) and weight matrices. Here, we found five MHC ligands in coat protein as 64-LSSFNGLGV-72; 86- RILEEDLIY-94; 96-MVFSITPSY-104; 100- ITPSYAGTF-108; 110- LTDDVTTED-118; having binding affinity and C terminal cleavage affinity is more than 0.5, which shows the surface activity of crystal structure of alfalfa mosaic virus RNA 3’UTR in complex with coat protein N terminal peptide. The predicted binding affinity is normalized by the 1% fractil. The MHC peptide binding is predicted using neural networks trained on C terminals of known epitopes. In analysis predicted MHC/peptide binding is a log transformed value related to the IC50 values in nM units. Total numbers of peptides found are 213.

![Surface activity of crystal structure of alfalfa mosaic virus RNA 3’UTR in complex with coat protein N terminal peptide, ABCD Chains](image)

Fig. 4.46- Surface activity of crystal structure of alfalfa mosaic virus RNA 3’UTR in complex with coat protein N terminal peptide, ABCD Chains
Table 4.4- SVM Based MHC-peptide allele binding nonamers in coat protein sequence

<table>
<thead>
<tr>
<th>ALLELE</th>
<th>Rank</th>
<th>Sequence</th>
<th>Residue No.</th>
<th>Peptide Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-Ab</td>
<td>1</td>
<td>PALKVPVVK</td>
<td>34</td>
<td>1.761</td>
</tr>
<tr>
<td>I-Ab</td>
<td>2</td>
<td>FCLTDDVTT</td>
<td>108</td>
<td>1.065</td>
</tr>
<tr>
<td>I-Ab</td>
<td>3</td>
<td>GKAGKPTKR</td>
<td>10</td>
<td>0.906</td>
</tr>
<tr>
<td>I-Ab</td>
<td>4</td>
<td>PVVKPTNTI</td>
<td>39</td>
<td>0.871</td>
</tr>
<tr>
<td>I-Ad</td>
<td>1</td>
<td>FGFELVFITA</td>
<td>142</td>
<td>0.677</td>
</tr>
<tr>
<td>I-Ad</td>
<td>2</td>
<td>HSYAVALCL</td>
<td>163</td>
<td>0.628</td>
</tr>
<tr>
<td>I-Ad</td>
<td>3</td>
<td>VWQSLGTPL</td>
<td>54</td>
<td>0.615</td>
</tr>
<tr>
<td>I-Ad</td>
<td>4</td>
<td>LSLSSFNGL</td>
<td>62</td>
<td>0.612</td>
</tr>
<tr>
<td>I-Ag7</td>
<td>1</td>
<td>QNYAALRKA</td>
<td>20</td>
<td>1.608</td>
</tr>
<tr>
<td>I-Ag7</td>
<td>2</td>
<td>KDFAGPRIL</td>
<td>80</td>
<td>1.509</td>
</tr>
<tr>
<td>I-Ag7</td>
<td>3</td>
<td>LDFDAOPEG</td>
<td>171</td>
<td>1.457</td>
</tr>
<tr>
<td>I-Ag7</td>
<td>4</td>
<td>PHGAFHANE</td>
<td>132</td>
<td>1.450</td>
</tr>
<tr>
<td>RT1.B</td>
<td>1</td>
<td>TFCLTDDVT</td>
<td>107</td>
<td>0.736</td>
</tr>
<tr>
<td>RT1.B</td>
<td>2</td>
<td>TAPTHAGMQ</td>
<td>149</td>
<td>0.725</td>
</tr>
<tr>
<td>RT1.B</td>
<td>3</td>
<td>SYAGTFCLT</td>
<td>103</td>
<td>0.617</td>
</tr>
<tr>
<td>RT1.B</td>
<td>4</td>
<td>PTKRSQNYA</td>
<td>15</td>
<td>0.568</td>
</tr>
</tbody>
</table>

4.6.2 ANN Based Prediction

CTLPred is a direct method for prediction of CTL epitopes crucial in subunit vaccine design. In direct methods the information or patterns of T cell epitopes instead of MHC binders were used for the development of methods. The method is based on elegant machine learning techniques like an Artificial Neural network and support vector machine. The methods also allow the consensus and combined prediction based on these two approaches. The artificial neural networks are crude electronic model based on the structure of brain. A network can consist of a few to a few billion neurons connected in an array of different methods. ANNs attempt to model these biological structures both in architecture and operation. The basic computational element (model neuron) of neural network is often called a node or unit. It receives input from some other units, or perhaps from an external source. Each input has an associated weight w, which can be modified. The unit computes some function f of the weighted sum of its inputs [69-72].

\[ \gamma_i = f\left(\sum_j \omega_{ij} \gamma_j\right) \]
Its output, in turn, can serve as input to other units. The weighted sum is called the net input to unit \( i \), often written \( \text{net}_i \). Note that \( w_{ij} \) refers to the weight from unit \( j \) to unit \( i \) (not the other way around). The function \( f \) is the unit's activation function. In the simplest case, \( f \) is the identity function, and the unit's output is just its net input. This is called a linear unit.

![Diagram of a neuron in an artificial neural network](image)

- **Basic Architecture:-**

The basic architecture of artificial neural network is shown through figure below

![Diagram of a neural network](image)

Fig. 4.47- The basic architecture of artificial neural network is shown through figure

There are many types of networks ranging from simple networks (Perceptrons) to complex self-organising networks (Kohonen networks). Similarly, there are many different kinds of learning rules used by neural networks, the common being the delta rule. The delta rule is often utilized by the most common class of ANNs called 'back
propagational neural networks' (BPNNs). Back propagation is an abbreviation for the backwards propagation of error. With the delta rule, as with other types of back propagation, 'learning' is a supervised process that occurs with each cycle or 'epoch' (i.e. each time the network is presented with a new input pattern) through a forward activation flow of outputs, and the backwards error propagation of weight adjustments. More simply, when a neural network is initially presented with a pattern it makes a random 'guess' as to what it might be. It then sees how far its answer was from the actual one and makes an appropriate adjustment to its connection weights. The other learning rule that is mostly used is feed-forward type of neural network [73, 74].

4.6.3 PSSM MHC-peptide binding predictions

**PSSMs for the prediction of MHC-peptide binding**

For a profile to be a good descriptor of the binding motif, peptides must be aligned by structural and sequence similarity. MHCI and MHCII molecules bind peptides in similar yet different modes and alignments of MHCI- and MHCII-ligands are obtained to be consistent with the binding mode of the peptides to their MHC class. Peptides that bind to a given MHC molecule share sequence similarity. Sequence patterns are traditionally used for the prediction of peptides binding to MHC molecules. Such sequence patterns are too simple, as the complexity of the binding motif cannot be precisely represented by the few residues present in the pattern. To overcome this limitation, we used RANKPEP Position Specific Scoring Matrices (PSSMs) or profiles from set aligned peptides known to bind to a given MHC molecule as the predictor of MHC-peptide binding [75].

**Information**- MHCI ligands are of short length (8-11), as they are constrained into the MHCI peptide binding groove, with their N- and C-terminal ends connected by a network of hydrogen bonds to conserved residues of the MHCI molecule. Thus, peptides bound to the same MHCI can differ by one or two amino acids, and as discussed, proper structural alignment of these peptides is better guaranteed if the peptides are of the same length. Accordingly, we have separated the peptides bound to a given MHCI molecule into subsets containing only peptides of the same length, and created separate PSSMs from ungapped block alignments. The peptide binding
groove of MHCII molecules is open, binding peptides in a manner that both the N- and C-terminus can extend beyond the binding groove, and thus, peptides bound to MHCII molecules display a great variability in length (9-22). Yet only a peptide core of 9 residues fits into the MHCII binding groove providing the binding energy. Poor amino acid sequence similarity between MHCII ligands together with their great variability in sequence length make them difficult to align. Thus, for the alignment of the MHCII ligands, we have used the motif discovery program MEME, including a priori information consistent with the MHCII-peptide binding mode: (A) there is only one binding core per MHCII ligand; (B) All the peptide sequences define the same motif and (C) the length of the motif is 9 [76-78].

**Interpretation of Results**

SVM classifiers trained on the sequence [1], physico-chemical properties [2] or on the combination of both [1+2] are able to distinguish the immunogenic and non-immunogenic peptides but with low accuracy. The classifier based on physico-chemical properties is able to outperform the classifier based on sequence alone, thus suggesting that the properties of peptides may be playing a critical role in deciding immunodominance. SVM-based immunodominance classifiers trained on residue properties and amino acid sequence were able to discriminate the immunogenic peptides from non-immunogenic peptides with an accuracy of 60.0% threshold of 0.5. This low accuracy may be due to the fact that we are pooling together all peptides regardless of MHCI restriction. Unfortunately, at this time there is not enough data on immunogenic and non-immunogenic for a single MHCI restriction element.
### Table 4.5 - Peptide binders to MHCI molecules of coat protein sequence

<table>
<thead>
<tr>
<th>MHCI-Peptide binders</th>
<th>*POS.</th>
<th>N</th>
<th>SEQUENCE</th>
<th>C</th>
<th>MW (Da)</th>
<th>SCORE</th>
<th>% OPT.</th>
</tr>
</thead>
<tbody>
<tr>
<td>11mer_H2_Db</td>
<td>17</td>
<td>KPT</td>
<td>KRSQNYAALRK</td>
<td>AQL</td>
<td>1316.53</td>
<td>86.0</td>
<td>40.57%</td>
</tr>
<tr>
<td>11mer_H2_Db</td>
<td>100</td>
<td>VFS</td>
<td>IPSYAGTFCCL</td>
<td>TDD</td>
<td>1154.35</td>
<td>84.0</td>
<td>39.62%</td>
</tr>
<tr>
<td>11mer_H2_Db</td>
<td>161</td>
<td>NQN</td>
<td>FKHSYAVACL</td>
<td>DFD</td>
<td>1233.5</td>
<td>77.0</td>
<td>36.32%</td>
</tr>
<tr>
<td>11mer_H2_Db</td>
<td>64</td>
<td>PLS</td>
<td>LSSFNGLGVRF</td>
<td>LYS</td>
<td>1178.36</td>
<td>76.0</td>
<td>35.85%</td>
</tr>
<tr>
<td>11mer_H2_Db</td>
<td>159</td>
<td>MQN</td>
<td>QNFKHSYAVAL</td>
<td>CLD</td>
<td>1259.43</td>
<td>71.0</td>
<td>34.99%</td>
</tr>
<tr>
<td>10mer_H2_Db</td>
<td>17</td>
<td>KPT</td>
<td>KRSQNYAALRK</td>
<td>KAQ</td>
<td>1188.36</td>
<td>77.0</td>
<td>34.84%</td>
</tr>
<tr>
<td>10mer_H2_Db</td>
<td>178</td>
<td>AQP</td>
<td>EGSKNPYSRF</td>
<td>NEV</td>
<td>1166.27</td>
<td>68.0</td>
<td>30.77%</td>
</tr>
<tr>
<td>10mer_H2_Db</td>
<td>209</td>
<td>LIT</td>
<td>VGLLDEADDL</td>
<td>DRH</td>
<td>1041.13</td>
<td>64.0</td>
<td>28.96%</td>
</tr>
<tr>
<td>10mer_H2_Db</td>
<td>156</td>
<td>HAG</td>
<td>MQQNFKHSY</td>
<td>AYA</td>
<td>1278.4</td>
<td>63.0</td>
<td>28.51%</td>
</tr>
<tr>
<td>9mer_H2_Db</td>
<td>17</td>
<td>KPT</td>
<td>KRSQNYAALRK</td>
<td>RKA</td>
<td>1032.17</td>
<td>102.0</td>
<td>69.86%</td>
</tr>
<tr>
<td>9mer_H2_Db</td>
<td>156</td>
<td>HAG</td>
<td>MQQNFKHSY</td>
<td>YAV</td>
<td>1115.22</td>
<td>78.0</td>
<td>53.42%</td>
</tr>
<tr>
<td>9mer_H2_Db</td>
<td>62</td>
<td>GTL</td>
<td>LSSLSSFNGL</td>
<td>GVR</td>
<td>919.05</td>
<td>63.0</td>
<td>43.15%</td>
</tr>
<tr>
<td>9mer_H2_Db</td>
<td>154</td>
<td>PTH</td>
<td>AGMQNQFK</td>
<td>HSY</td>
<td>1019.13</td>
<td>57.0</td>
<td>39.04%</td>
</tr>
<tr>
<td>9mer_H2_Db</td>
<td>28</td>
<td>LRG</td>
<td>AQLPDPPLAL</td>
<td>KVP</td>
<td>916.14</td>
<td>57.0</td>
<td>39.04%</td>
</tr>
<tr>
<td>8mer_H2_Db</td>
<td>65</td>
<td>LSL</td>
<td>SFINGNSGF</td>
<td>RFL</td>
<td>761.83</td>
<td>99.0</td>
<td>55.31%</td>
</tr>
<tr>
<td>8mer_H2_Db</td>
<td>115</td>
<td>DDV</td>
<td>TTEDGRAV</td>
<td>AHG</td>
<td>829.86</td>
<td>98.0</td>
<td>54.75%</td>
</tr>
<tr>
<td>8mer_H2_Db</td>
<td>18</td>
<td>PTK</td>
<td>RQNYAAL</td>
<td>RKA</td>
<td>904.04</td>
<td>81.0</td>
<td>45.25%</td>
</tr>
<tr>
<td>8mer_H2_Db</td>
<td>57</td>
<td>VVQ</td>
<td>SLGTPLSL</td>
<td>SSF</td>
<td>768.91</td>
<td>79.0</td>
<td>44.13%</td>
</tr>
<tr>
<td>8mer_H2_Db</td>
<td>202</td>
<td>RAG</td>
<td>PLRSLTv</td>
<td>GLL</td>
<td>880.1</td>
<td>76.0</td>
<td>42.46%</td>
</tr>
</tbody>
</table>

*POS-Amino acids position in sequence, N-Terminal-N, C-Terminal-C, MW-Molecular Weight, %OPT-Optimal Score for given MHCI peptide binder in Mouse.

### Table 4.6 - Peptide binders to MHCII molecules of coat protein sequence

<table>
<thead>
<tr>
<th>MHCII-Peptide binders</th>
<th>*POS.</th>
<th>N</th>
<th>SEQUENCE</th>
<th>C</th>
<th>MW (Da)</th>
<th>SCORE</th>
<th>% OPT.</th>
</tr>
</thead>
<tbody>
<tr>
<td>I_Ab.p</td>
<td>94</td>
<td>DLI</td>
<td>YRMVFSITP</td>
<td>SYA</td>
<td>1095.33</td>
<td>13.779</td>
<td>38.67%</td>
</tr>
<tr>
<td>I_Ab.p</td>
<td>120</td>
<td>EDG</td>
<td>RAVAHGPM</td>
<td>QEF</td>
<td>934.08</td>
<td>12.843</td>
<td>36.04%</td>
</tr>
<tr>
<td>I_Ab.p</td>
<td>148</td>
<td>ELV</td>
<td>FTAPTHAGM</td>
<td>QNO</td>
<td>914.04</td>
<td>11.612</td>
<td>32.59%</td>
</tr>
<tr>
<td>I_Ab.p</td>
<td>197</td>
<td>RKA</td>
<td>FPRAGPLRSL</td>
<td>LIT</td>
<td>982.17</td>
<td>10.747</td>
<td>30.16%</td>
</tr>
<tr>
<td>I_Ab.p</td>
<td>165</td>
<td>KHS</td>
<td>YAVALCLDF</td>
<td>DAQ</td>
<td>996.2</td>
<td>10.614</td>
<td>29.79%</td>
</tr>
<tr>
<td>I_Ag7.p</td>
<td>121</td>
<td>DGR</td>
<td>AVAHGPMQ</td>
<td>EFP</td>
<td>906.02</td>
<td>13.89</td>
<td>33.98%</td>
</tr>
<tr>
<td>I_Ag7.p</td>
<td>166</td>
<td>HSY</td>
<td>AVAVALCDFD</td>
<td>AQP</td>
<td>948.11</td>
<td>13.46</td>
<td>32.93%</td>
</tr>
<tr>
<td>I_Ag7.p</td>
<td>87</td>
<td>GPR</td>
<td>ILEEDLIYR</td>
<td>MVF</td>
<td>1145.34</td>
<td>12.426</td>
<td>30.40%</td>
</tr>
<tr>
<td>I_Ag7.p</td>
<td>21</td>
<td>RSQ</td>
<td>NYYAALRKa</td>
<td>LPK</td>
<td>1016.17</td>
<td>10.164</td>
<td>24.87%</td>
</tr>
<tr>
<td>I_Ag7.p</td>
<td>198</td>
<td>KAF</td>
<td>PRAGPLRSL</td>
<td>ITV</td>
<td>948.15</td>
<td>10.085</td>
<td>24.67%</td>
</tr>
<tr>
<td>I_Ad.p</td>
<td>118</td>
<td>TTE</td>
<td>DGRAVAHGN</td>
<td>PMQ</td>
<td>877.91</td>
<td>12.73</td>
<td>23.95%</td>
</tr>
<tr>
<td>I_Ad.p</td>
<td>96</td>
<td>IYR</td>
<td>MVFSITPSY</td>
<td>AGF</td>
<td>1026.22</td>
<td>12.173</td>
<td>22.91%</td>
</tr>
<tr>
<td>I_Ad.p</td>
<td>161</td>
<td>NQN</td>
<td>FKHSHAVCL</td>
<td>CLD</td>
<td>1017.2</td>
<td>10.357</td>
<td>19.49%</td>
</tr>
<tr>
<td>I_Ad.p</td>
<td>68</td>
<td>SSF</td>
<td>NGLGVRFLY</td>
<td>SFL</td>
<td>1020.2</td>
<td>7.366</td>
<td>13.86%</td>
</tr>
</tbody>
</table>

*POS-Amino acids position in sequence, N-Terminal-N, C-Terminal-C, MW-Molecular Weight, %OPT-Optimal Score for given MHC II peptide binder in Mouse.

The output consists of a list of peptides ordered by their binding potential (score) to the selected MHC molecule.

- **RANK**: Relative rank of the predicted peptide
- **POS**: Position of the peptide in the input protein
- **N**: Amino acid sequence of the three residues preceding the N-terminus of the predicted peptide
- **SEQUENCE**: Amino acid sequence of the predicted peptide
- **C**: Amino acid sequence of the three residues following the C-terminus of the predicted peptide
- **MW(Da)**: Molecular weight in Daltons of the predicted peptide
SCORE- Score of the peptide
% OPT.- Percentile score of the predicted peptide relative to that of the consensus.
The consensus is the sequence that yields the maximum score, namely optimal score, with the selected profile.

Models
In Rankpep we have chosen an immunodominance filter based on SVM-based classifier trained on both residue properties and amino acid sequence. If this filter is set ON only those peptides that are classified as immunodominant are returned by the server. Since immunodominance classification is threshold dependent we have given three optional thresholds to choose from.

Position-specific scoring matrix or profile
Profiles basically consist of a table listing the observed sequence-weighted frequency of all amino acids in every column of a sequence alignment. Peptide alignments and PSSMs for the prediction of MHCI and MHCII were obtained differently.

It includes a selection of 102 and 80 PSSMs for the prediction of peptide binding MHCI and MHCII molecules, respectively. Several PSSMs for the prediction of peptide binders of different sizes are usually available for each MHCI molecule. By default, we use PSSMs for the prediction of peptides of 9 residues. MHCII-specific PSSMs are always for the prediction of peptide binders of 9 residues.

Sequence input
Input for analysis is either be protein sequence or FASTA format
MSSSQKKGKAGKPKTKRSQNYAALRKAQLPKPPALKVPVKPTNITLPQTG

Calculation
Sequence variability is calculated from multiple amino acid sequence alignments as indicated, using variability metric (V) formally identical to the Shannon entropy equation as:
Where Pi is the fractions of residues of amino acid type i, and M is equal to 20, the number of amino acid types. V ranges from 0 (total conservation, only one amino acid type is present at that position) to 4.322 (all 20 amino acids are equally represented in that position). Note that in order to achieve the maximum value V = 4.3, at least 20 sequences are required. Gap symbols (-) are considered for deriving the consensus sequence but are not computed for the variability calculations. Given a sequence variability threshold Vt, the consensus sequence is generated from the sequence alignment only for those positions with a V \leq Vt as the most common amino acid, whereas variable position positions (V > Vt) are masked and represented in the consensus sequence with a "." symbol. Segments with a position masked are not considered in the predictions of MHC-peptide binding.

4.7 Comparative Modeling
A novel method was developed for fold recognition/homology modeling, in which a large sequence database is iteratively searched to construct a sequence profile until a template can be found in a database of proteins with known structure. The method differs from the PDB-BLAST method in that a sequence profile is only made if a template is not readily found in the database of known structures. A sequence profile is subsequently made for the template, using the same number of PSI-BLAST iterations that were used to identify it. Query and template sequences are subsequently aligned using a score based on profile-profile comparisons. The alignment score is modified so as to ensure unreliable parts of the alignment is discarded. Neural networking was developed for fold recognition/homology modeling, in which a large sequence database is iteratively searched to construct a sequence profile until a template can be found in a database of proteins with known structure. CPHModel reveals information on many aspects of protein structure and function, such as protein interaction expression pattern, surface activity, binding sites, and electrostatic potentials. The corresponding atoms derived from the alignment are extracted from the template file and used as a starting point for the homology modeling [79].
Molecular modeling is primarily a tool for calculating the energy of a given molecular structure. Thus, the first step in designing a molecular modeling investigation is to define the problem as one involving a structure-energy relationship.

Fig. 4.48- Comparative Modeling of Cucumber mosaic virus focused on computational approach to deciphering the sequence similarity, molecular modeling and their function of coat protein of cucumber mosaic virus.
4.8 Automatic modeling of protein three-dimensional structure

The structure generated from these Automatic comparative molecular modeling methods allows identification of potential drug targets. We used Geno3D is an automatic web server for protein molecular modeling. Starting with a query protein sequence, the server performs the homology modeling in six successive steps: (i) identify homologous proteins with known 3D structures by using PSI-BLAST; (ii) provide the user all potential templates through a very convenient user interface for target selection; (iii) perform the alignment of both query and subject sequences; (iv) extract geometrical restraints (dihedral angles and distances) for corresponding atoms between the query and the template; (v) perform the 3D construction of the protein by using a distance geometry approach and (vi) finally send the results to the user. These regions are antigenic in nature and form antibodies against plant diseases and it should be tried on gel separation to get a pure form of it for primer prediction. The target structure also serves as a detailed model for determining the structure of peptide within that protein structure [80].

**Automatic Modeling - Papaya ringspot virus**

Significant alignments of protein sequence of RNA-directed RNA polymerase (Papaya ringspot virus) show identify homologous proteins with known 3D structures by using PSI-BLAST i.e. it shows 75 % similarity with polyprotein from Japanese yam mosaic virus.

The electrostatic potential is the energy of interaction of a point positive charge (an electrophile) with the nuclei and electrons of a molecule. Negative electrostatic potentials indicate areas that are prone to electrophilic attack. For example, a negative electrostatic potential of benzene (left) shows that electrophilic attack should occur onto the p system, above and below the plane of the ring, while the corresponding electrostatic potential for pyridine (right) shows that an electrophile should attack the nitrogen in the s plane, and not the p system of the ring. A sufficiently small value of the electron density provides overall molecular size and shape (as given by a conventional space-filling or CPK model). The electrostatic potential can then be mapped onto the electron density by using color to represent the value of the potential. The resulting model simultaneously displays molecular size and shape and electrostatic potential value. For example, the electrostatic potential of benzene can be
mapped onto the electron density. Colors toward “red” indicate negative values of the electrostatic potential, while colors toward “blue” indicate positive values of the potential.

For example, the electrostatic potential map of the zwitterionic form of b-alanine shows the negative carboxylate (red) and the positive ammonium (blue) termini separated by the neutral (green) carbon chain.
Fig. 4.49-Electrostatic potentials displaying the distribution of the electric charge at the molecular surface allows studying protein-protein or protein-substrate interactions. An electrostatic potential map conveys information about the distribution of charge in a molecule and structure shows the electron rich areas (Blues) and electron deficient areas (Red).

Fig. 4.50-Ribbon structure of coat protein
**Fig. 4.51 - Ramachandran plot coat protein structure**

![Ramachandran Plot](image)

### Plot statistics

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residues in most favored regions (A, LB)</td>
<td>153</td>
</tr>
<tr>
<td>Residues in additionally allowed regions (AL)</td>
<td>71</td>
</tr>
<tr>
<td>Residues in generously allowed regions (G)</td>
<td>17</td>
</tr>
<tr>
<td>Residues in disallowed regions</td>
<td>7</td>
</tr>
<tr>
<td>Number of non-glycine and non-polar residues</td>
<td>246</td>
</tr>
<tr>
<td>Number of end residues (excl. Cty and Pro)</td>
<td>2</td>
</tr>
<tr>
<td>Number of glycine residues (shown as triangles)</td>
<td>17</td>
</tr>
<tr>
<td>Number of proline residues</td>
<td>22</td>
</tr>
<tr>
<td>Total number of residues</td>
<td>777</td>
</tr>
</tbody>
</table>

*Based on analysis of 18 Peptide structures of resolution 2.0 or better and 5.25% of residues having more than 90% in the most favored regions.*
4.9 Molecular Modeling

Bioinformatics software tools reveal information on many aspects of protein structure and function, such as protein interaction expression pattern, surface activity, binding sites, and electrostatic potentials. The study is focused on computational approach for deciphering the sequence similarity, molecular modeling and their function of coat protein. The data generated from these assay is subsequently added to a searchable database that allows potential drug targets to identify. Molecular modeling is a collection of (computer based) techniques for deriving, representing and manipulating the structures and reactions of molecules, and those properties that are dependent on these three dimensional structures. This aims to introduce in a simple way the hierarchy of computational modeling methods used nowadays as standard tools by organic chemists for searching for, rationalising and predicting structure and reactivity of organic, bio-organic and organometallic molecules. The emphasis will be on helping to develop a feel for the correct "tool" to use in the context of a typical problem in structure, activity or reactivity, by describing the limitations and strengths of each method [81, 82].

Modeling Modeling - Cucumber Mosaic Virus (CMV)

There are Eight antigenic determinants site in coat protein sequence were seen

<table>
<thead>
<tr>
<th>n</th>
<th>Start Position</th>
<th>Sequence</th>
<th>End Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>ANFRVLSQQLS</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>YTFTSITLKP</td>
<td>77</td>
</tr>
<tr>
<td>3</td>
<td>86</td>
<td>YGKRLLPDSVTE</td>
<td>98</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>DKKLVSRLQIRVNPILPKFDSTVW</td>
<td>141</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VTVRKVPASSDLHSVAAISA</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>145</td>
<td>DGASPVLVYQYAASG</td>
<td>159</td>
</tr>
<tr>
<td>6</td>
<td>162</td>
<td>ANNKLLYDL</td>
<td>170</td>
</tr>
<tr>
<td>7</td>
<td>181</td>
<td>RKYAVLVYSK– Beta-Sheet</td>
<td>190</td>
</tr>
<tr>
<td>8</td>
<td>195</td>
<td>ETDELVLHVDEIEH – Beta-Sheet</td>
<td>207</td>
</tr>
</tbody>
</table>
Fig. 4.52- Comparative Modeling of Tomato aspermy virus capsid protein shows nine antigenic determinants sites

Fig. 4.53- Structure of antigenic peptide having sequence – 30-ANFRVLSQQLS-40 of viral coat protein from Cucumber Mosaic Virus
Fig. 4.54- Structure of antigenic peptide having sequence –68-YTFTSITLKP-77 of viral coat protein from Cucumber Mosaic Virus

Fig. 4.55- Structure of antigenic peptide having sequence – 86-YGKRLLLDPDSVTE-98 of viral coat protein from Cucumber Mosaic Virus
Fig. 4.56- Structure of antigenic peptide having sequence – 100-DKKLVSRLQIRVNPLPKFDSTVWVTVRKVPASSDLSVAAISA-141 of viral coat protein from Cucumber Mosaic Virus.

Fig. 4.57- Structure of antigenic peptide having sequence – 145-DGASPVLVYQYAASG-159 of viral coat protein from Cucumber Mosaic Virus.
Fig. 4.58 - Structure of antigenic peptide having sequence – 162- ANNKLLYDL-170 of viral coat protein from Cucumber Mosaic Virus

Fig. 4.59 - Structure of antigenic peptide having sequence –181- RKYAVLVYSK-190 of viral coat protein from Cucumber Mosaic Virus

Fig. 4.60 - Structure of antigenic peptide having sequence –195- ETDELVLHVDIEH-207 of viral coat protein from Cucumber Mosaic Virus
4.10 Protein analysis

Blocks represent a conserved region in the MSA. Blocks differ from profiles in lacking insertion and deletion positions in the sequences. Blocks are made by searching for a section of MSA alignment that is highly conserved. However, aligned regions are also found by searching each sequence in the turn for similar patterns of same length. These patterns include a region with one or few matching character followed by a short spacer region until the sequences start to be different. These patterns are all of the same length and they are aligned, the matching sequence characters will appear in columns. The alignments of this type are performed by searching patterns in sequences. Several Blocks located in different regions in set of sequences are used to produce MSA, and Blocks are constructed from a set of aligned sequence pairs. Statistical and Bayesian statistical methods are also used to locate the most alike regions of sequences [83, 87].

Fig. 4.61- Phylogenetic analysis of Potyviridae plant viruses’ family shows the Blocks, which represent a conserved region in the MSA. Blocks differ from profiles in lacking insert and delete positions in the sequences.

Fig. 4.62- Phylogenetic analysis of Potyviridae plant viruses’ family shows the motif regions developed by GIBBS, which represent a conserved region in the MSA. A
GIBBS analysis differs from profiles in lacking insert and deletes positions in the sequences.

Since approximately the same two blocks are reported using both MOTIF and GIBBS and include all thirty polyprotein sequences submitted, it is very likely that these blocks represent correct alignments. Indeed, Lawrence et al [88] indicate that these alignments are identical to those determined from analysis of the 3-dimensional structures of these proteins, and that these 2 regions are the only ones in common for the group. Notice, however, that MOTIF apparently aligned potyviridae plant virus species incorrectly in the Block. Please note that, because the GIBBS algorithm is non-deterministic, (using randomly determined starting points) the GIBBS results may differ when the same sequences are submitted repeatedly. Although Block Maker will always use the same seed for the random number generator and always sorts the sequences alphabetically by name, a small change to the sequences (even changing a sequence name and thus its order in the input presented to GIBBS) can change the results.

**BLOCK Analysis interpretation**

- Blocks represent a conserved region in the MSA. Blocks differ from profiles in lacking insert and delete positions in the sequences.
- Blocks may be made by searching for a section of MSA alignment that is highly conserved.
- However aligned regions may also be found by searching each sequence in the turn for similar patterns of same length.
- These patterns may include a region with one or few matching character followed by a short spacer region until the sequences start to be different. These patterns are all of the same length and they are aligned, the matching sequence characters will appear in columns.
- The first alignments of this type were performed by computer program search patterns in sequences.
- Several Blocks located in different regions in set of sequences may be used to produced MSA, and Blocks may be constructed from a set of align sequence pairs.
• Statistical and Bayesian statistical methods are also to locate the most alike regions of sequences.

**Block summary**

• Given the huge variety of methods for computing phylogenies, how can the biologist determine what is the best method for analyzing a given data set?
• Here we take thirty plant virus species and evolutionary analysis of a family of polyprotein sequences were determined by phylogenetic analysis.
• The evolutionary relationships among the polyprotein sequences were depicted by placing the sequences as outer branches on a tree.
• The evolutionary mechanisms giving rise to changes in polyprotein sequences and the challenges faced when aligning protein sequences are discussed.
• Blocks were constructed from a set of align maturase protein sequence pairs.
• Motif analyses of thirty plant virus species polyprotein sequences are representing five blocks- a conserved region in the multiple sequence alignment.
• However aligned regions may also be found by searching each sequence in the turn for similar patterns of same length having PSSM width are 47.
• GIBBS analyses of thirty polyprotein sequences are representing four blocks- a conserved region having PSSM width are 47.
• These patterns include a region with highest 47 matching character followed by a short spacer region until the sequences start to be different.
• These patterns are all of the same length and they are aligned, the matching sequence characters will appear in columns.
• Phylogenetics often makes use of numerical data, (numerical taxonomy) which can be scores for various “character states” such as the size of a visible structure or it can be DNA sequences.
• Similarities and differences between organisms can be coded as a set of characters, each with two or more alternative character states.
• In an alignment of DNA sequences, each position is a separate character, with four possible character states, the four nucleotides.
Conclusion

- These patterns include a region with highest 47 matching character followed by a short spacer region until the sequences start to be different.
- These patterns are all of the same length and they are aligned, the matching sequence characters will appear in columns.
- This work provides insight into the evolutionary history for the gene families of the potyviridae plant virus species, linking the expansion of these families to the duplications of the gene cluster regions, and showing that they are composed of subgroups with distinct evolutionary (and possibly functional) differences.
- Such regions represent conserved functional or structural domains in predicting the function and structure of polyprotein from thirty plant virus species.
- In identifying new members of protein families.
- This work provides insight into the evolutionary history for the polyproteins of thirty plant virus species, linking the expansion of these strains to the duplications of the gene cluster regions.
- Showing that they are composed of subgroups with distinct evolutionary (and possibly functional) differences.
- Such regions represent conserved functional or structural domains in predicting the function and structure of polyproteins from thirty plant virus species and in identifying new members of protein families.

Protein analysis - Groundnut ringspot virus

- Analysis of Nucleocapsid protein
- Number of atoms - 4103
- Molecular Formula- C1278H2085N333O393S14
- Molecular Weight -28852.545
- Number of Residues- 258
- Percent Hydrophilic amino acids – 52.3256
- Percent Hydrophobic amino acids – 47.6744
γ Ratio of % OF Hydrophilic to % Hydrophobic – 1.09756
γ Mean Beta Hydrophobic moment -0.21467
γ Mean Helix Hydrophobic moment-0.19133
γ Number of Basic amino acids-36
γ Number of Acidic amino acids-34
γ Estimated pI for Protein-8.6
γ Total Linear Linear Charge Density -0.27907
γ Polar Area of Extended Chain (Angs) – 17257.5
γ Non Polar Area of Extended Chain (Angs) –28904.6
γ Total Area of Extended Chain (Angs) –46162.1
γ Analysis of Nucleocapsid protein
γ Polar ASA of folded protein (Angs) -4073.35
γ Non Polar ASA of folded protein (Angs) -7252.92
γ ASA of folded protein (Angs) -11326.3
γ Ratio of Folded of protein to extended area- 0.265277
γ Buried polar area of Folded of protein (Angs) -10970.4
γ Buried Non polar area of Folded of protein (Angs) -19119.8
γ Buried Charge area of Folded of protein (Angs)-1253.76
γ Total buried surface (Angs)-31344.0
γ Number of buried amino acids -83
γ Packing volume (est) (Angs) -34734.2
γ Packing volume (act) (Angs) -34529.8
γ Interior volume of protein –24749.0
γ Exterior volume protein – 9780.83
γ Partial specific volume (Ml/g)-0.734116
γ Fisher volume ratio (act)- 0.395202
γ Protein solubility – 1.61211
γ Solvent free energy of folding (Kcal/mol) = -239.4
γ Total number of negatively charged residues (Asp + Glu): 34
γ Total number of positively charged residues (Arg + Lys): 36
Protein analysis - Cowpea mosaic virus (CPMV)

γ Analysis of Coat protein
γ Percent Hydrophilic amino acids – 49.7132
γ Percent Hydrophobic amino acids – 50.2868
γ Ratio of % OF Hydrophilic to % Hydrophobic – 0.988593
γ Mean Beta Hydrophobic moment -0.207361
γ Mean Helix Hydrophobic moment-0.168268
γ Number of Basic amino acids-105
γ Number of Acidic amino acids-97
γ Estimated pI for Protein-8.9
γ Total Linear Linear Charge Density -0.195029
γ Polar Area of Extended Chain (Angs) – 66160.3
γ Non Polar Area of Extended Chain (Angs) –117985.0
γ Total Area of Extended Chain (Angs) –184145.0
γ Polar ASA of folded protein (Angs) -13198.0
γ Non Polar ASA of folded protein (Angs) -17611.9
γ ASA of folded protein (Angs) -30809.8
γ Analysis of Coat protein
γ ASA of folded protein (Angs) -30809.8
γ Ratio of Folded of protein to extended area- 0.179129
γ Buried polar area of Folded of protein (Angs) -49208.9
γ Buried Non polar area of Folded of protein (Angs) -85764.2
γ Buried Charge area of Folded of protein (Angs)-5632.88
γ Total buried surface (Angs)-140597.0
γ Number of buried amino acids -541
γ Packing volume (est) (Angs) -141932.0
γ Packing volume (act) (Angs) -138932.0
γ Interior volume of protein –102030.0
γ Exterior volume protein – 36901.4
γ Partial specific volume (ML/g)-0.726493
γ Fisher volume ratio (act)- 0.361671
γ Protein solubility – 1.38252
Solvent free energy of folding (Kcal/mol) = -1019.52
Total number of negatively charged residues (Asp + Glu): 97
Total number of positively charged residues (Arg + Lys): 105

4.11 Summary

Multiple sequence analysis of *geminiviruses* assumed to be an estimate of a phylogeny; branch lengths are proportional to the amount of inferred evolutionary change (Fig. 4.1). Transport of the viral genome into the nucleus is an obligatory step in the replication cycle of plant geminiviruses. In these virus types, the multifunctional coat protein (CP) of *Indian mungbean yellow mosaic virus* (MYMV) is thought to be involved in this process (Fig. 4.2a to 4.2c). Given a set of gene sequences, it should be possible to reconstruct the evolutionary relationships among genes and among organisms. The evolutionary mechanisms giving rise to changes in coat protein sequences and the challenges faced when farmers are getting different pesticide response to the viral infections.

Phylogenetics assumed to be an estimate of a phylogeny; branch lengths are proportional to the amount of inferred evolutionary change in potyviruses and Tombusviridae (Table 4.1 to 4.2). Blocks were constructed from a set of align protein sequence pairs (Fig. 4.3 to 4.10). Motif analyses of thirty sequences are representing five blocks, which shows the distinct amino acids and some region are identical which are conserved from these species which having the phylogenetic resemble from the same family. These patterns include a region with highest 47 matching character followed by a short spacer region until the sequences start to be different (Fig. 4.59 to 4.60). These patterns are all of the same length and they are aligned, the matching sequence characters will appear in columns.

The study is focused on computational approaches for deciphering the antigenic epitope and their function of coat protein from plant viruses. Fragment identified through this approach tend to be high-efficiency binders, which is a larger percentage of their atoms are directly involved in binding as compared to larger molecules. The data generated from these assay is subsequently added to a searchable database that allows potential drug targets to identify.
Gomase and Kale (2007), Hopp and Woods antigenicity, Welling antigenicity, Parker antigenicity, Protrusion Index antigenicity, B-EpiPred Server antigenicity, Kolaskar and Tongaonkar antigenicity, Emini surface activity, Karplus-Schulz flexibility (fig. 4.11 to 4.21) antigenicity scales were designed to predict the locations of antigenic determinants. The Robson-Garnier and SOPMA method predicted the secondary structure of coat protein (fig. 4.22 to 4.24). This methods show high antigenic regions present in beta sheet response than helical region of this peptide.

We also found the Sweet hydrophobicity, Kyte & Doolittle hydrophobicity, Abraham & Leo hydrophobicity, Bull & Breese hydrophobicity, Guy hydrophobicity, Miyazawa hydrophobicity, Roseman hydrophobicity, Cowan HPLC pH7.5 hydrophobicity, Rose hydrophobicity, Eisenberg hydrophobicity, Manavalan hydrophobicity, Black hydrophobicity, Fauchere hydrophobicity, Janin hydrophobicity, Rao & Argos hydrophobicity, Wolfenden hydrophobicity, Wilson HPLC hydrophobicity, von Heijne Hydrophilicity, Chothia hydrophobicity scales (fig. 4.25 to 4.46). These scales are essentially a hydrophilic index, with apolar residues assigned negative values.

A coat protein sequence from *alfalfa mosaic virus* is 221 residues long, having antigenic MHC binding peptides. MHC molecules are cell surface glycoproteins, which take active part in host immune reactions and involvement of MHC class-I and MHC II in response to almost all antigens (Table 4.4 to 4.6). Bepipred predicted epitopes antigenicity determinant shows epitopes present in the AMV eliciting the desired immune response (Table 4.3). Epitopes are the sites of molecules that are recognized by antibodies of the immune system in mouse and rabbit. Knowledge of Epitopes may be used in the design of vaccines and diagnostics tests. It is therefore of interest to develop improved methods for predicting epitopes. In this analysis, we describe an improved method for predicting linear epitopes. The region of maximal hydrophilicity is likely to be an antigenic site, having hydrophobic characteristics, because terminal regions of coat proteins is solvent accessible and unstructured, antibodies against those regions are also likely to recognize the native protein. It was shown that a coat protein is hydrophobic in nature and contains segments of low complexity and high-predicted flexibility. Predicted antigenic fragments can bind to MHC molecule is the first bottlenecks in vaccine design.
Conclusion

Peptide fragments from plant viruses involved multiple antigenic components to direct and empower the immune system to protect the host from infection. MHC molecules are cell surface proteins, which take active part in host immune reactions and involvement of MHC class-I & II in response to almost all antigens. Predicted MHC binding regions acts like red flags for antigen specific and generate immune response against the parent antigen. So a small fragment of antigen can induce immune response against whole antigen. This theme is implemented in designing subunit and synthetic peptide vaccines. The sequence analysis method is allows potential drug targets to identify active sites which form resistance against plant diseases. The method integrates prediction of peptide MHC class I binding; proteosomal C terminal cleavage and TAP transport efficiency. Antigenic epitopes of coat protein are important antigenic determinants against the viral attack.

4.12 References


