1. Introduction

Bioinformatics

Bioinformatics is a discipline which originally arose for the utilitarian purpose of introducing order into the massive data sets produced by the new technologies of molecular biology. These techniques originated with large-scale DNA sequencing and the need for tools for sequence assembly and for sequence annotation, i.e., determination of locations of protein-coding regions in DNA. A parallel development was the construction of sequence repositories.

Bioinformatics is the mathematical, statistical and computing methods that aim to solve biological problems using DNA, RNA and proteins sequences and related information. Bioinformatics is about searching biological databases, comparing sequences looking at protein sequence for computational simulation and more generally, solving biological question with a computer. Biological or genetic information is a fundamental concept of bioinformatics. This is growing exponentially with a shift in emphasis from individual biomolecules, to analysis of how they interact in complex networks which control the developmental and physiological processes of whole biological systems, and research into how this relates to human health. This transition has increased the importance of bioinformatics and raises key challenges which make it imperative that computer scientists work closely with biologists to refine existing bioinformatics tools and develop new ones.

The ability to sort and extricate genetic code from a viral genomic database of millions base pair of DNA in a meaning full way is perhaps a simplest form of bioinformatics. Moving on to another level, bioinformatics is useful in mapping different viral's, genomes and deriving differences in their genetic makeup through computational simulation. What is more complex is to decipher the genetic code itself to see what the difference in genetic makeup between different peoples translates into in terms of physiological traits.

Computational simulation of experimental biology is another application for bioinformatics which is apply to refer to as 'in silico' testing. This area will expand in prolific way, given a need to obtain a greater degree of predictability in animal and human clinical trials. Added to this, is interesting scope that in silico' testing provides to deal with the brewing hostility towards animal testing. Sequence based searching is

another key skill for biologist. Identifying homologous sequences provides bases for phylogenetic analysis and sequence pattern recognition [**1**].

Role of computer science in molecular biology

Among the tools of applied mathematics some are of special importance, namely probability theory and statistics and algorithms in computer science. A large amount of research in bioinformatics uses and combines methods from these two areas. Computer-science algorithms form the technical background for bioinformatics, in the sense that the operation and maintenance of bioinformatics databases require the most up-to-date algorithmic tools. Probability and statistics, besides being a tool for research, also provides a language for formulating results in bioinformatics. In a very short interval 'computational biology' has become an extremely active research field. Although it began with sequence analysis, it now encompasses a far wider spread of activity, which truly epitomizes modern scientific research;

(a) It is highly interdisciplinary, requiring at least computational, mathematical, biological, physical and chemical knowledge;
(b) Its implementation may furthermore require knowledge of computer science, chemical engineering, biotechnology, medicine, pharmacology, etc.; and
(c) There is little distinction between the work carried out in the public domain, either in academic institutions (universities) or state research laboratories, or privately by commercial firms.

The handling and analysis of DNA, RNA and protein sequences remains one of the prime tasks of bioinformatics. This topic is usually divided into two parts, functional genomics and proteomics, which seeks to determine the role of the sequence in the living cell, either as a transcribed and translated unit (i.e. a protein) or as a regulatory motif, etc., and comparative genomics and proteomics, in which the sequences from different organisms, or even different individuals, are compared in order to determine ancestries and correlations with disease. Clearly the sequence analysis of unknown sequences with known ones can also help to elucidate functions; both parts are concerned with the finding of patterns or regularities which is the core of all scientific work. One can feel that it is fortunate (for scientist) that life is in some sense encapsulated in such a highly formalized object as a sequence of symbols.

The requirement of entire genomes and proteomes to feel this search has led to tremendous advances in the technology for rapid sequencing, which in turn has put

new demands on informatics for interpreting the raw output of a sequencer. If a DNA sequence is the message, then functional genomics is already concerned with meaning of the message, and in turn this has led to the experimental analysis of the RNA transcript (transcriptome) and the repertoire of expressed proteins (proteome), each of which presents fresh informatics challenges [**2, 3**].

Proteomics and its application

Current major research and clinical applications of bioinformatics include its use to improve the diagnosis and detection of diseases, to promote vaccine development by screening databases for pathogen proteomes and genomes, and to increase the understanding of evolutionary processes through analysis of nucleotide/protein sequence mutations.

Proteomics consists of the mathematical, statistical and computing methods that aim to solve biological problems using amino acid sequences and related information. Proteomics is about searching biological databases, comparing sequences, looking at protein structure and more generally, asking biological question with a computer. Major difficulties for target identification include the structural analysis of proteins and their detection. These technologies are compared to enable the selection of the one by matching the needs of a particular project. There are prospects for further improvement, and proteomics technologies will form an important addition to the existing genomic and chemical technologies for new target validation. Biological or genetic information concepts provide a distinct knowledge layer for biologists, especially when they become interested in high throughput experimental analyses though computational methods are the fundamental concept of proteomics. Proteomics technologies have produced target sites of drug targets, which are creating a bottleneck in drug development process. There is an increasing need for better target validation for new drug development and proteomic technologies are contributing to it. Identifying a potential protein drug target within a cell is a major challenge in modern drug discovery; techniques for screening and analysis the proteomes are very easier through computational simulations. Proteomics is applicable for protein analysis and bioinformatics based analysis gives the comprehensive molecular description of the actual protein component. Bioinformatics is being increasingly used to support target validation by providing functionally predictive information mined

from databases and experimental datasets using a variety of computational tools [**4, 5**].

Plant Virus:

Viruses are very small (submicroscopic) infectious particles (virions) composed of a protein coat and a nucleic acid core. They carry genetic information encoded in their nucleic acid, which typically specifies two or more proteins.

Plant viruses, like all other viruses, are obligate intracellular parasites that do not have the molecular machinery to replicate without the host. The plant viruses are defined as viruses pathogenic to higher plants. Viruses are very small and can only be seen under an electron microscope. The structure of a virus is given by its coat of proteins, which surround the viral genome. Assembly of viral particles takes place spontaneously. Over 50% of known plant viruses are rod shaped (flexious or rigid). Exact length is normally dependent on the genome but it is usually between 300–500 nm with a diameter of 15–20 nm. Protein subunits can be placed around the circumference of a circle to form a disc. In the presence of the viral genome, the discs are stacked, and then a tube is created with room for the nucleic acid genome in the middle. The second most common structure amongst plant viruses are isometric particles. They are 40–50 nm in diameter. In cases when there is only a single coat protein, the basic structure consists of 60 T subunits, where T is an integer. Some viruses may have 2 coat proteins are the formation of the particle is analogous to a football.

A very small number of plant viruses have, in addition to their coat proteins, a lipid envelope. This is derived from the plant cell membrane as the virus particle buds off from the cell. Viruses are very small (submicroscopic) infectious particles (virions) composed of a protein coat and a nucleic acid core. They carry genetic information encoded in their nucleic acid, which typically specifies two or more proteins. Plant viruses are not nearly as well understood as the animal counterparts; Plant viruses are grouped into 73 genera and 49 families. Plant viruses are similar to animal viruses in most basic characteristics but they can also be markedly different. Most plant viruses have RNA as the genetic material.

Inadequate sequences and data are available on plant viruses, which need further expansion and revision. These sequence need to be characterized for their use in agribiotech and pharmaceuticals. Thus, bio-computational tools are useful for high throughput screening and proteomics analysis to correlate large and diverse datasets.

Proteomics research requires analysis of large sequences and is heavily dependent upon bioinformatics computation where strict standards, requiring careful data selection for protein model building, followed by adequate testing and validation through bioinformatics tools required [**6-8**].

Computational Vaccine Design

Computational Drug Design is dealing with the mathematical modeling of flexibility of biomolecules and its application in the field of Virtual Screening. In all cases, the aim of using the computer for drug design is to analyze the interactions between the drug and its receptor site and to "design" molecules that give an optimal fit. The techniques provided by computational methods include computer graphics for visualization and the methodology of theoretical chemistry. The structure of the drug molecule that can specifically interact with the biomolecules can be modeled using computational tools; by means of quantum mechanics the structure of small molecules can be predicted to experimental accuracy. Statistical mechanics permits molecular motion and solvent effects to be incorporated [**9**].

Vaccines have mostly been composed of killed or attenuated whole pathogens. For safety reasons, however, it could be desirable to use peptide vaccines that are able to generate an immune response against a given pathogen. Such vaccines could contain peptides representing linear epitopes from the proteins of the pathogen. Linear epitopes induce protective immunity in mice against host antigen. By immunizing animals, synthetic peptides containing linear epitopes can also be used to raise antibodies against a specific protein, which e.g. can be used in screening assays or as diagnostic tools. Epitopes are parts of proteins or other molecules that antibodies (made by B-cells) bind. Most protein epitopes are composed of different parts of the polypeptide chain that are brought into spatial proximity by the folding of the protein, identification of such linear peptide segments will often be the initial step in the search for antigenic determinants in pathogenic organisms. The traditional experimental peptide scanning approach is clearly not feasible on a genomic scale. Prediction methods are very cost effective and reliable methods for predicting linear B-cell epitopes would therefore be a first step in guiding a genome wide search for B-cell antigens in pathogenic organism [**10-12**].

The new paradigm in vaccine design is emerging, following essential discoveries in immunology and development of new major histocompatibility complex (MHC)

Class binding peptides prediction tools. MHC molecules are cell surface glycoproteins, which take active part in host 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 our work designing subunit and synthetic peptide vaccines. In this study we report on the binding ability of antigenic peptides to MHC class, which integrated prediction of peptide MHC class binding; proteasomal C terminal cleavage and TAP transport efficiency of protein [**13-15**].

2. Motivation of research

A. System architecture in plant virus's database

Bioinformatics databases contain massive amounts of experimental data. Browsing and analyzing these data is fascinating and will surely lead to many interesting discoveries. The developing projects concerned with searching for interesting information in bioinformatics databases belong to the most vital area in scientific research. A large volume of information on the structure and expression regulation of various plant viruses has been accumulated. The Kyoto Encyclopedia of Genes and Genomes (KEGG) represent a database consisting of known genes and their respective biochemical functionalities. KEGG is a knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information. The Pathway database is supplemented by a set of ortholog group tables for the information about conserved subpathways (pathway motifs), which are often encoded by positionally coupled genes on the chromosome and which are especially useful in predicting gene functions. [**16**]. Database contains more information about pathway maps, pathway modules, functional hierarchies and ontologies to metabolic pathways. However diseases, symptoms and host species of plant viruses are not given.

A database cryoEM database is vital for keeping track of the very large number of images collected and analyzed by the automated system and essential for quantitatively evaluating the utility of methods and algorithms used in the data collection. The database can be accessed using a variety of tools including specially developed Web-based interfaces that enable a user to annotate and categorize images

using a Web-based form [**17**]. Limitation is that the only images and information are given. Taxonomical classification and reference are not given

DPVweb provides a central source of information about viruses, viroids and satellites of plants, fungi and protozoa. Comprehensive taxonomic information and classified lists of virus sequences are provided. The database also holds information for all sequences of viruses, viroids and satellites of plants, fungi and protozoa. As far as possible, nomenclature for genes and proteins are standardized within genera and families [**18**]. Collections of plant viruses, symptoms are less as the available dispersed data. Also specially coat protein or capsid protein with references are not included, which are important for infection on host plan species. Many new viral entries are not mentioned; also cannot search by plant virus name, Accession Number and standard taxonomical identification number.

Secondary database construction is an important subject in the field of bioinformatics. As the full genomic sequences of some organisms are being completed and followed by structural and functional studies, construction of secondary database becomes essential on the agenda. The rice dwarf virus (RDV) is a pathogen infecting rice in China, Japan and the Southeastern Asia region and leading to considerable economic loss. Based on the data generated from recent genomic research and earlier biochemical studies scattered in various primary databases and scientific journals, we have constructed a compact, user-friendly and non-redundant job-oriented secondary database [**19**]. Only specific area wise plant virus data are collected and important of rice dwarf virus is taken for consideration. Search techniques are not seen.

Genome Information Broker for Viruses (GIB-V) is a comprehensive virus genome/segment database. We extracted 18 418 complete virus genomes/segments from the International Nucleotide Sequence Database Collaboration (INSDC) by DNA Data Bank of Japan (DDBJ), EMBL and GenBank and stored them in our system. The list of registered viruses is arranged hierarchically according to taxonomy [**20**]. Only classifications of viral genomes are given. Species information, sequential data are available.

CoVDB provides a convenient platform for rapid and accurate batch sequence retrieval, the cornerstone and bottleneck for comparative gene or genome analysis of annotated coronavirus genes and genomes. Sequences can be directly downloaded from the website in FASTA format. CoVDB also provides detailed annotation of all coronavirus sequences using a standardized nomenclature system and overcomes the

problems of duplicated and identical sequences in other databases [**21**]. Only coronavirus data are collected and taken for consideration. Search techniques are not seen.

354 viruses were originally published in paper form by the Association of Applied Biologists (AAB) between 1970 and 1989, while additional descriptions have been added to a CD-ROM (also published by the AAB) since 1998 [**22**]. Only virus name are given with their family names, here lack coat protein data and references are not added. Also only 354 viruses are added.

B. Development of customized Biotool for gene prediction

Statistical or *ab initio* methods attempt to predict genes based on statistical properties of the given DNA sequence. The simplest way to detect potential coding regions is to look at Open Reading Frames (ORFs). An ORF is a sequence of codons in DNA that starts with a Start codon (ATG), ends with a Stop codon (TAA, TAG or TGA) and has no other (in-frame) stop codons inside [**23, 24**]. Tools take a given DNA sequence and search within each of the possible reading frames stop codons and its corresponding start codon. Output is specific for the particular family of species only.

C. Computational Modeling and Simulation for drug design

Despite there is not infallible method to predict antigenic peptides, there are several rules that can be followed to determine the peptide fragments from a protein that are likely to be antigenic in nature. These rules are also dictated to increase the odds of an Ab recognizing the native protein.

1) Antigenic peptides should be located in solvent accessible regions and contain both hydrophobic and hydrophilic residues.
    a) For proteins of known 3D structure solvent accessibility can be determined using a variety of programs such as DSSP, NACESS or WHATIF, among others. A web server to calculate solvent accessibility using Whatif [**25, 26**]. Limitation to the study of viral coat proteins is that very few protein having crystal structure data. Also we need Unix-based system (SGI, SUN, HP, Linux, etc.) and Fortran 77 compiler.
    b) If the 3D structure is not known, use any of the following web servers to predict accessibilities: PHD, PredictProtein, NNPREDICT, NSSP, JPRED, PredAcc (c), ACCpro [**27, 28**]. The accuracy for predicting either the alpha-

helix or beta-strand in proteins with higher alpha-helix or beta-strand content, respectively, should be greatly improved. This is multiple alignment-based neural network system. Accuracy of prediction is below 70%.

2) Preferably select peptides lying in long loops connecting Secondary Structure (SS) motifs, avoiding peptides located in helical regions. This will increase the odds that the Ab recognizes the native protein.
   a) For protein with known 3D coordinates, SS can be obtained from the sequence link of the relevant entry at the Brookhaven data bank. The PDBsum server also offers SS analysis of pdb records. Very limited plant virus structural data available, not applicable for protein sequences.
   b) When no structure is available secondary structure predictions can be obtained from any of the following servers: PSI-PRED, NNSP [**29**]. Using a very stringent cross validation method to evaluate the method's performance, PSIPRED-2.6 achieves an average Q3 score of 80.7%. But useful for higher organism having structural data.
3) When possible, choose peptides that are in the N- and C-terminal region of the protein. Because the N- and C- terminal regions of proteins are usually solvent accessible and unstructured, Abs against those regions are also likely to recognize the native protein.
4) For vaccine design, different tools are used for essential discoveries in immunology and prediction of new major histocompatibility complex (MHC) Class binding peptides [**30,31**].
5) Tools for the detection of antigenic peptides: Several methods based on various physio-chemical properties of experimental determined epitopes (flexibility, hydrophobicity, accessibility) have published for the prediction of antigenic determinants. Perhaps the simplest method for the prediction of antigenic determinants is that of Kolaskar and Tongaonkar, which is based on the occurrence of amino acid residues in experimentally determined epitopes [**32**]. Predicted epitopes length and position are varying for different tolls and results are not seen in one specific tool. Findings are not accurate as the standard antigen antibody reaction.

3. Limitation to the earlier research contribution

The research survey brought the following limitations in the solutions proposed by earlier researchers

- Plant virus database having more information about pathway maps, pathway modules, functional hierarchies and ontologies to metabolic pathways. However diseases, symptoms and host species data are missing.
- Only specific area wise plant virus data are collected and few species of viruses are taken into consideration.
- There is no separate database available for the plant viruses in India.
- The database can be accessed specially developed Web-based interfaces that enable to a normal user to annotate and categorize the viral data. Limitation is that the only images and information are given. Taxonomical classification and reference for plant viral data are not mention with virus names.
- Collection of plant viruses, symptoms are less as the available dispersed data. Also specially coat protein or capsid protein with references are not included, which are important for infection on host plant species. Many new viral entries are not mention; also cannot search by plant virus name, Accession Number and standard taxonomical identification number.
- Database standard are missing in some databases.
- Tools are taking input as a given DNA sequence and search within each of the possible reading frames stop codons and its corresponding start codon. Output is specific for the particular family of species only not all organisms.
- Limitation for the study of plant viral coat proteins is that very few proteins having 3D crystal structure data. Also we need higher end servers as Unix-based system (SGI, SUN, HP, Linux, etc.) and Fortran 77% to analyzed them. Simple protocols are not available for researchers regarding the physicochemical parameter study of plant viruses.
- The accuracy for predicting either the alpha-helix or beta-strand in proteins with higher alpha-helix or beta-strand content, respectively, should be greatly improved.
- Accuracy of prediction of sequence analysis tools is below 70 %. Using a very stringent cross validation method to evaluate the method's performance, and useful for higher organism having structural data.

- Predicted epitopes length and position are varying for different tools and results are not seen in one specific tool. Findings are not accurate as the standard antigen antibody reaction.
- The performance of prediction method has not measured through threshold dependent parameters such as sensitivity, specificity, NPV, PPV and accuracy.
- Plant viral peptides research are not favors of position mostly possesses higher volume, aromatic, hydrophobic and accessible residues; also not more focus on the N and C terminal of the peptides prefers the higher volume, charged, aromatic, hydrophobic and accessible residues.

The factors that contribute to the limitations

(1) The success of structural genomics and proteomics initiatives requires the new annotated database development and application of software's for structural analysis, prediction, and annotation.

(2) Till 2001 various datasets of plant viruses are in patenting format

(3) A well-defined sequence pattern is observed it is generally an indication of an evolutionary relationship, suggestive of a functional relationship, rather than resulting from the sequence requirements for a particular fold

(4) Very less researchers are bridging the gap of the comprehensive selection of biological data analysis and computational simulation tools alongside an advanced query system and a context-mapping tool that implements a relevancy model towards new target finding from various data sources.

(5) Many prediction methodology and tools are available for analysis of hydrophobicity and accessible residues, MHC and epitopes from the given sequential data. Also results are varying from every tool.

(6) Techniques used for study of biological system and out put data from bioinformatics tools are not easy to interpret for new target validation.

(7) The physical and chemical nature of the proteins offers an approach to the analysis of function, for the same every time we have to analyse the physiochemical parameters of given proteins [**33**].

4. Objective of the project

This research focus on the system architecture development for plant viruses and development, application of software's for protein structural analysis, prediction, and annotation for bioinformatics based drug designing. Following are the objective of the research

- Development of database and high throughput screening of Plant viruses.
- Sequence analysis of Plant viruses and Gene prediction.
- Comparative genomics and proteomics analysis.
- Creation of customized Database architectures including database access control.
- Development of customized software Biotools for database of plant viruses.

Solutions to address the objective

This research contributes the following solutions to address the above issues:

- Collection of all plant viruses' data, symptoms from high throughput screening of the available dispersed data. Also specially coat protein or capsid protein with references are included, which are important for infection on host plant species. Many new viral entries are also added; we can search by plant virus name, accession number and standard taxonomical identification number.
- These databases prepared describing Accession Number of Plant Virus, GenBank Identification Number of Proteins from Virus, Taxonomy ID, Species Name, Protein Name, Taxonomy (Classification of Plant Virus), FASTA of coat Protein of Plant Virus, Reference for protein sequences and virus data.
- Comprehensive viruses customized database architectures of all plants including database access control also developed.

- *ab initio* methods are used to predict genes based on statistical properties of the given DNA sequence. Tool can predict the complementary DNA, ORFs from any DNA sequence and search within each of the possible reading frames stop codons and its corresponding start codon.
- We analysed the different plant viral species using various bioinformatics servers for comparative genomics and proteomics analysis of viruses. 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 viral sequences are representing blocks- a conserved region in the multiple sequence alignment.
- 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 antigenicity prediction methods. Predictions are based on plots that reflect the occurrence of amino acid residues in experimentally known segmental epitopes.
- Above theme is implemented in designing subunit and synthetic peptide vaccines from plant viruses. Antigenic epitopes of coat protein are important antigenic determinants against the viral attack. The sequence analysis method is allows potential drug targets to identify active sites which form resistance against plant diseases.

5. Work Plan

5.1 Phase I

Proteomics system architecture of plant virus's

This section contains the detailed descriptions of individual plant viruses. About 1837 individual descriptions of plant viruses are now included. A databases allows a search to be made using a text query on the complete set of descriptions and this can be limited to selected fields (subheadings in the descriptions) if required. Within each virus description, there are links to access the all information and to display references in a small, separate, frame at the new page. Identification, screening, classification and browsing of protein sequences of plant viruses were carried out. The program therefore provided a comprehensive resource on plant viruses and virus diseases for use in research, quarantine, extension and education.

This eventually led to the development of a database product that incorporated all the original descriptions with some new ones and much extra information on plant virus classification and protein sequences. All the plant viruses information was originally issued imported into a computer database and displayed API. This will allowed users to search and display the descriptions more easily than the old paper format. The original format of the descriptions was kept as much as possible, but some new references were introduced so that relevant data could be accessed more easily. The text of these descriptions was not changed, except to correct typographical errors and to complete references to papers in press at the time of publication. Usually, new plant virus information was also provided. Features we are added as:

- An additional descriptions of plant viruses that had not previously been published
- An up-to-date taxonomic treatment of plant viruses, satellites and viroids, showing all the recognized families, genera and species and with brief family and genus descriptions.
- Lists of all plant virus, satellite and viroid sequences available from the research articles, public databases classified by their current species name (information not easily and reliably obtainable directly from the databases themselves).

- For about proteins sequences, additional annotation was provided together to display interactive viral diseases, from which the sequences of particular genes or other features (e.g. coat proteins) could be easily extracted for analysis.

The program therefore provided a comprehensive resource on plant viruses and virus diseases for use in research, quarantine, extension and education.

Plants virus's data were entered and annotated along with the flexibility to add customized annotation features to virus's entries. The data were exported regularly in the database architecture software tool format, with incrementally performing major revisions, recoding and updating data. A data structure can be viewed as an interface between two functions or as an implementation of methods to access storage that is organized according to the associated data type.

5.2 Chapter 2. Solution for system architecture in plant virus's database

High throughout put screening and multiple sequence analysis is required for molecular study of plant diseases in India. This will help in development of plants virus's database plants. This object model having optimal structure depends on the natural organization of the applications of virus data and on the applicant's requirements. Data structures include Accession Number of Plant Virus, GenBank Identification Number of Proteins from Virus, Taxonomy ID, Species Name, Protein Name, Taxonomy (Classification of Plant Virus), FASTA of coat Protein of Plant Virus, Reference for protein sequences and virus data. Comprehensive viruses customized database architectures of all plants including database access control, which optimized to deal with very large amounts of viral data stored on a permanent data storage device. It also controls the security of the database. Data security prevents unauthorized users from viewing or updating the database. Organizations of coat protein and plant virus database use one kind of DBMS for daily viral transaction processing. This will helpful for researcher to compare their datasets and such information for genetic research on plant viruses. Overall systems design decisions are performed by viral data and there are common computations requested on attributes such as counting, summing, averaging, sorting, grouping, cross-referencing of plant viruses.

The factor we considered while proposing this solution are:

A. Coat protein database

This database provides a central source of information about plant viruses, viroids and satellites of plants, fungi and protozoa, with some additional data on related viruses. Following are the standard for Coat protein database

- *Accession Number of Plant Virus*
- *GenBank Identification Number of Proteins from Virus*
- *Taxonomy ID*
- *Species Name- Nomenclature*
- *Protein Name*
- *Taxonomy (Classification of Plant Virus*
- *FASTA of coat Protein of Plant Virus*
- *Reference*

B. Plant virus database

Plant virus database provides a central source of information about plant viruses, viroids and satellites of plants, fungi and protozoa, with some additional data on related viruses. Following are the standard for plant virus database

- *Virus Name*
- *TAXID*
- *Classification*
- *Symptoms*
- *Host Species*
- *References*

Features of system architecture

Input Data

A set of plant viruses with known biological response forms the basis to performed study. A convenient worksheet environment allows easy visualization, selection and manipulation of the viruses' related data. The activity data can be imported/inserted from text; commas separated and excel files and then MS Access in the worksheet.

Viral classification

Virus classification involves naming and placing viruses into a taxonomic system. Like the relatively consistent classification systems seen for cellular organisms, virus classification is the subject of ongoing proposals. This is largely due to the pseudo-living nature of viruses, which are not yet definitively living or non-living. As such, they do not fit neatly into the established biological classification system in place for cellular organisms, such as plants and animals, for several reasons. Virus classification is based mainly on phenotypic characteristics, including morphology, nucleic acid type, mode of replication, host organisms, and the type of disease they cause. A combination of two main schemes is currently in widespread use for the classification of viruses. We used Baltimore classification system, which places viruses into one of seven groups. These groups are designated by Roman numerals and separate viruses based on their mode of replication, and genome type. Accompanying this broad method of classification are specific naming conventions and further classification guidelines set out by the International Committee on Taxonomy of Viruses.

Viruses can be placed in one of the seven following groups:

a. I: dsDNA viruses
b. II: ssDNA viruses (+)sense DNA
c. III: dsRNA viruses
d. IV: (+)ssRNA viruses (+)sense RNA
e. V: (-)ssRNA viruses (-)sense RNA
f. VI: ssRNA-RT viruses (+)sense RNA with DNA intermediate in life-cycle
g. VII: dsDNA-RT viruses

Preliminary Data Analysis

Preliminary data analysis utility allows preliminary investigation of data with the help of univariate statistical analysis (minimum, maximum, average, variance, and so on), cross-correlation to visualize data patterns and inter relationships.

Data Preprocessing

Data preprocessing utility allows effective reduction of the data by removing the invariable columns. Data preprocessing is performed by various scaling methods. In addition, a variance cut-off option is provided to filter off the noise in the data.

Data Selection

A data selection feature allows selection of dependent and independent variables to be considered in the study. Training and test sets can be created by manual, random and sphere exclusion methods.

5.3 Phase II

ORF prediction tools

With the rapid accumulation of genomic sequence information, there is a pressing need to use computational approaches to accurately predict gene structure. Computational gene prediction is a prerequisite for detailed functional annotation of genes and genomes. The process includes detection of the location of open reading frames (ORFs) and delineation of the structures of introns as well as exons if the genes of interest are of eukaryotic origin. The ultimate goal is to describe all the genes computationally with near 100% accuracy. The ability to accurately predict genes can significantly reduce the amount of experimental verification work required. Genes are the functional elements of the genome and represent an important goal in any mapping or sequencing project. This section involves the following function:

- This section aims to develop program which can detect open reading frames in single gene; whereas others may seek to identify every gene in the genome. This depends on the size and complexity of the genome, and the availability of genetic and physical maps.
- This is defined as a series of sense codons beginning with an initiation codon (usually ATG) and ending with a termination codon (TAA, TAG or TGA). The simplest way to detect a long ORF is to carry out a three-frame translation of a query sequence using a program such as ORF Finder.
- The scope of the work, contemporary gene-finding in plant viruses generally involves the scanning of raw sequence data for long open reading frames.

- Plant viruses' genomes are small and the genes lack introns. Problems may be encountered identifying small genes, genes with unusual organization or genes using rare variations of the genetic code.

Gene prediction using Markov Models and Hidden Markov Models

Markov models and HMMs can be very helpful in providing finer statistical description of a gene. A Markov model describes the probability of the distribution of nucleotides in a DNA sequence, in which the conditional probability of a particular sequence position depends on previous positions.

The use of Markov models in gene finding exploits the fact that oligonucleotide distributions in the coding regions are different from those for the noncoding regions. These can be represented with various orders of Markov models. Since a fixed-order Markov chain describes the probability of a particular nucleotide that depends on previous nucleotides, the longer the oligomer unit, the more non randomness can be described for the coding region. Therefore, the higher the order of a Markov model, the more accurately it can predict a gene. Because a protein-encoding gene is composed of nucleotides in triplets as codons, more effective Markov models are built in sets of three nucleotides, describing nonrandom distributions of trimers or hexamers, and so on. The parameters of a Markov model have to be trained using a set of sequences with known gene locations. The parameters of the model are established, it can be used to compute the nonrandom distributions of trimers or hexamers in a new sequence to find regions that are compatible with the statistical profiles in the learning set.

Gene finding in Plant viruses' therefore involves more complex analysis, based on specific motifs (signals) differences in base composition compared to surrounding DNA (content) and relationships with known genes (homology). No gene prediction algorithm in eukaryotes is 100% reliable. Problems may include the failure to detect exons, the detection of phantom exons, mis-specification of exon boundaries and exon fusion.

- Fifth-order Hidden Markov Models to look at nucleotide frequency and dependency data are useful for the identification of shadow genes.

- However, as the number of completely sequenced plant viruses' genomes increases, it is becoming a more common practice to find such genes by analyzing genomic sequences with known virus, using databank search.
- ORF Finder and similar programs also give the user a choice of standard or variant genetic codes, as minor variations are found among the prokaryotes and in mitochondria. The principles of searching for genes in mitochondrial and chloroplast genomes are much the same as in other organisms [**35-39**].

Role of Computer Science

Today's mapping and sequencing projects generate enormous amounts of sequence data very rapidly. The manual annotation methods are available; while highly discriminatory and applicable to unsequenced DNA fragments, do not have the throughput to cope with the amount of data produced. Bioinformatics methods are necessary in large-scale sequencing projects to temporarily fill the information gap between sequence and function. A number of algorithms have been developed to scan and annotate raw sequence data, but it should be appreciated that none of the methods discussed are 100% reliable and any computer based gene predictions should be confirmed using other experimental methods.

5.4 Chapter 3- Development of customized Biotool for gene prediction

Statistical and ab initio methods are used to predict genes based on statistical properties of the given DNA sequence. Tool can predict the complementary DNA, ORFs from any DNA sequence and search within each of the possible reading frames stop codons and its corresponding start codon.

ORF prediction

The simplest way to detect potential coding regions is to look at Open Reading Frames (ORFs). An ORF is a sequence of codons in DNA that starts with a Start codon (ATG), ends with a Stop codon (TAA, TAG or TGA) and has no other (in-frame) stop codons inside.

Model Building Method

A variety of model building methods ranging from linear like MLR, PCR, PLSR to non-linear like neural networks, k-nearest neighbor are available in this utility. The

resulting Fifth-order Hidden Markov Models are shown in terms of descriptors involved in the model and their coefficients. The models can be compared based on corresponding statistical parameters and contribution/fitness plots.

Evaluate lengths of ORFs, an algorithm would then take a given DNA sequence and search within each of the possible reading frames stop codons and its corresponding start codon. For each such potential ORF determine the length and evaluate that.

Evaluate codon usage:

Here we use the fact that codon usage in coding regions differs substantially from that in non-coding regions. A number of these measures have been proposed, such as codon usage or hexamer counts. The codon usage of a string of DNA is given by a 64-component vector that counts how many times each codon is present in the string. The in-phase hexamer feature measures the frequency of occurrence of oligonucleotides of length six in a specific reading frame. Hexamer counts are mostly modeled as fifth-order Hidden Markov Models.

Fifth-order: $P(X_n = s| \cap_j <_n X_j) = P(X_n = s|Xn-_1X_n-_2X_n-_3X_n-_4X_n-5)$

Model Validation

The qualities of models obtained are compared based on internal (cross-validation) and external validation (predicted) parameters as well as standard error. In addition, randomization test is also performed to judge quality of ORF prediction with respect to nucleotide sequences in model.

Prediction

The developed model can be utilized to predict activity of newly designed molecules. The developed ORF prediction model can be coupled with nucleotide sets generation utility as a tool for virtual screening.

Gene tool consist of menus as File, Edit, View, Analyze on the given sequence. The sequences used in the analysis are available in the example folder so we can perform each operation. Gene Tool is helpful for those scientists who want a quick result before research and it also helpful for ORF prediction in genetic engineering and molecular biology. The user can compare the original sequence and the input sequence. Also the DNA sequence can be analyzed for later reference. Gene tool

deals with DNA sequence assembly; gene finding and analysis, protein expression and regulation. Gene predicts examine some of the application of computational biology. Gene predicts allows researchers to precisely adopt genes and gene products to suite their specific requirement. Gene Tool also includes the various operations as reverse sequence, double stranded sequence, complement of sequence, features Information that helps the researcher to find out the function of particular gene sequence.

5.5 Phase III

Computational Modeling and Simulation for Drug Design

Drug discovery can be thought of as the work done from the time of the identification of a therapeutic need in a particular disease area to the time the drug candidate deemed most likely to safely affect the desired therapeutic benefit is identified. This drug candidate may be a small molecule or a biological macromolecule such as a protein or nucleic acid. Drug discovery activities vary between small molecules and biological macromolecules, but, once a drug candidate has been identified and moves into the drug development phase, the regulatory governance of nonclinical and clinical research and the marketing approval process is very similar in both cases.

It is generally recognized that drug discovery and development are very time and resources consuming processes. There is an ever growing effort to apply computational power to the combined chemical and biological space in order to streamline drug discovery, design, development and optimization. In biomedical arena, computer-aided or in silico design is being utilized to expedite and facilitate hit identification, hit-to-lead selection, optimize the absorption, distribution, metabolism, excretion and toxicity profile and avoid safety issues. Commonly used computational approaches include Structure-based drug design (drug-target interaction), and quantitative structure-activity and quantitative structure-property relationships [**40-43**].

Synthetic peptide vaccines- Each protein antigen has one or more epitopes. These short amino acid sequences can be synthesized in a machine and it was suggested that the resulting peptides might be used as vaccines. Compared with traditional vaccines it would be easier to ensure the absence of contaminants such as viruses and proteins. In this virus there is an important epitope within the virion protein VP1. Synthetic

peptides of this sequence induced reasonable levels of neutralizing and protective antibodies in laboratory animals.

DNA vaccines- The most revolutionary approach to vaccination is the introduction into the vaccinee of DNA encoding an antigen, with the aim of inducing cells of the vaccinee to synthesize the antigen. One advantage of this approach is that there is a steady supply of new antigen to stimulate the immune system, as with live virus vaccines. Because the antigen (a virus protein in this case) is produced within the cells of the vaccinee, it is likely to stimulate efficient cell mediated responses [**44-48**].

A. Computer Analysis of Sequence Data

For the development of antigenic determinant sites we used following methodology are used for Computational Modeling and Simulation for Drug Design from plant viruses. Analysis can be done on sequences of roteins. Some of the applications under sequence analysis module include:

Sequence Alignment- Multiple alignments of protein sequences are important tools in studying sequences. The basic information they provide is the 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.

Phylogenetic tree construction methods and programs- We produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

In Silico prediction of antigenic epitopes– Prediction of antigenicity program predicts those segments from within viral coat protein that are likely to be antigenic by eliciting an antibody response. Antigenic epitopes is determined using the Gomase-Kale methods, Hopp and Woods, Welling, Parker and Protrusion Index (Thornton) antigenicity methods. Predictions are based on plots that reflect the occurrence of amino acid residues in experimentally known segmental epitopes.

Secondary structure prediction- The important concepts in secondary structure prediction by GOR and SOPMA method are identified as: 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.

Solvent accessible regions - We also show that there is an equal balance between the hydrophobic and the hydrophilic accessible surfaces of the 3D protein surfaces irrespective of the protein size. This results in a patchwork surface of hydrophobic and hydrophilic areas, which could be important for protein interactions and/or activity. Solvent-accessibility calculations have been applied to the protein folding problem, energy calculations and structure refinement. Atomic accessible areas in relation to distinguishing correctly and incorrectly folded protein structures. Hydrophobic surface area has been incorporated into energy minimization calculations. Finding the location in solvent accessible regions in protein, type of plot determines the hydrophobic and hydrophilic scales and it is utilized for prediction. This may be useful in predicting membrane-spanning domains, potential antigenic sites and regions that are likely exposed on the protein surface [**49-51**].

In Silico prediction of peptide–MHC binding affinity - The challenge of predicting which peptide sequences bind to which major histocompatibility complex (MHC) molecules has been met with various computational techniques. Scoring matrices, hidden Markov models, and artificial neural networks are examples of algorithms that have been successful in MHC–peptide-binding prediction. Because these algorithms are based on a limited amount of experimental peptide-binding data, prediction is only possible for a small fraction of the thousands of known MHC proteins. In the primary field of application for such algorithms—vaccine design—the ability to make predictions for the most frequent MHC alleles may be sufficient. However, emerging applications of leukemia-specific T cells require a patient-specific MHC–peptide-binding prediction. The modular model of MHC presented here is an attempt to maximize the number of predictable MHC alleles, based on a limited pool of experimentally determined peptide-binding data.

We also predicted peptide binders to MHCI and MHCII molecules from protein sequences or sequence alignments using Position Specific Scoring Matrices (PSSMs). For development of MHC binder, an elegant machine learning technique SVM has been used. This method will be useful in cellular immunology, Vaccine design,

immunodiagnostics, immunotherapeutics and molecular understanding of autoimmune susceptibility [**52-57**].

B. Structure based drug design

Development of computational tools that will improve effectiveness and efficiency of drug discovery and development process, decrease use of animals, and increase predictability. It is expected that the power of CADD will grow as the technology continues to evolve. Structure based drug design (SBDD) is a well established methodology built on the knowledge of three dimensional structure of target protein. We analysed the coat protein structure of plant viruses using following methodology-

Proteomics module provides structural analysis

- *Proteomics Analysis*- Comparative Proteomics and Genomics Sequence Analysis
- *Target Identification, Validation and Prediction*

3D Modeling

The main purpose of 3D Modeling and analysis module is to be able build, analyze and predict 3 dimensional structures of macromolecules and macromolecular complexes. In the 3D Modeling module comprises of the following features

- *Homology modeling* - Homology Modeling involves prediction of the 3-dimensional structure of a protein sequence based on sequential similarity (typically > 30 %) with a set of molecules (templates).
- *Threading (Energy Based)* - In which the energy of the conformation adopted by a protein is estimated. This is used to evaluate the structure adopted by a protein.
- Simulation of molecular systems using molecular dynamics, Monte Carlo methods, Genetic Algorithms and Simulated Annealing employing a variety of force-fields. Electrostatic calculations using the Poisson-Boltzmann equation are also incorporated.

Structural Analysis

Proteomics module provides structural analysis tools as below:

- *Surfaces and Volumes*- Any molecule interacts with the surrounding solvent molecules through its accessible area/volume. This module predicts the accessible and Molecular surface area/volume of a protein. The methods are 1. Shrake and Rupley; 2. Lee Richards; 3.Connolly [**58, 59**].
- *Binding Site Detection*- A drug molecule attaches itself to a protein molecule at a particular site. This site is known as the binding site. Determination of this site is the objective of this module. There are two algorithms in Bio-Suite ™ to predict binding sites
- *Interactions*- The atoms of a molecule have a tendency to interact with the neighboring atoms resulting in van der Waals, hydrophobic, hydrogen, salt bridge, aromatic-aromatic and amino-aromatic interactions are checked for in a given 3D molecule.
- *Quality Check*- Any protein's 3D structure has to conform to a set of established standards. These include parameters like the bond lengths, bond angles, planarity, chirality, disulphide bond check, Ramachandran plot and the Z factor. Deviations from known standard values are calculated and reported. It is based on Procheck.
- *Fold Classification* - Fold Classification functionality uses SSAP(Sequential Structure Alignment Program)algorithm. The Fold Classification functionality detects the 3-D fold for a given three-dimensional structure of target protein. It finds for the target protein the best fold from the fold database. It uses a double dynamic programming method to identify similar structural regions.

Computational Simulation

The "Simulations" module essentially simulates the behaviour of a molecular system on a computer. It provides a direct route from the microscopic details of a system to macroscopic properties of experimental interest. The different sub modules covered are as:

- ***Force field-*** The potential energy function (PEF) and the associated parameters constitute a force field. Typically, PEF is a sum of terms that correspond to bond stretch, bond angle, torsion angle, van der Waals and electrostatic interaction energies as functions of conformation. While the bond stretch, bond angle, torsion angle and improper torsion terms are summed over

all bonds, angles, torsions and improper torsions, the van der Waals and electrostatic terms are summed over all possible pairs of atoms. The option to use either AMBER or CHARMM force field for macromolecules and uses GAFF for small molecules. For each of the force fields, there is an option to choose the type of dielectric: either constant or distant dependent [**52-62**].

- ***Energy Minimization***- The process of energy minimization adjusts the coordinates of the molecular system so as to lower its energy, relative to the starting conformation. First order unconstrained minimization methods are among the most popular choice of energy minimization procedures adopted currently. This is due to various reasons, prominent among them being considerations of storage requirements and speed of convergence. Typically these methods require only information about the function value and its first derivative at each step.

Advantages of this methodology

- Alignment of diverse set of molecules finding common template can also be aligned based on a set of atoms selected in same order in all the molecules of the set.
- Allows alignment of given set of molecules in the protein active site with respect to the co-crystallized ligand.
- Quick and accurate editing of existing molecules.
- Comprehensive manipulations for visualization, chirality, charges, labeling and so on can be performed with ease.
- Readily calculate and display important molecular features such as atomic charges, atomic measurements and features such as molecular chirality.
- Optimizes the geometry of a given protein molecule and helps to attain local minimum energy structure.
- Allows for batch optimization of a set of molecules in addition to optimization of a single molecule.
- Provides additional features like definition of aggregates or constraints to the molecule during optimization. With aggregation, a few atoms or side chains or the entire backbone of the protein can be frozen to avoid change in position during optimization.

- Retaining the backbone structure is very much desired during optimization of a protein and hence this utility comes in handy for modeler aiming to optimize a protein structure without disturbing its back bone conformation.
- Allows automated homology modeling from the selected template. In addition, the user defined template and alignment can also be used for automated homology modeling.
- Allows modifying the alignment generated by automated homology modeling tool and using it for generating the model.
- Supports manual modeling by allowing the mutation, insertion, deletion and excision.
- We can insert amino acids of our choice based on the RMSD and similarity of the various hits obtained by modeling from its database.
- The depiction of protein cavities in terms of shape and ability to texture the cavity surface with molecular properties provide in depth understanding of cavities.
- The cavity shape enables choice of appropriate ligand and its conformation that is likely to fit in the cavity.
- The electrostatic properties mapping on cavity and channel surface allows identification of regions of charge localization that can guide ligand placement as well as comparison of cavities and channels.
- The coat protein sequence of plant viruses was analyzed to study the antigenicity, solvent accessible regions and MHC class peptide binding, which allows potential drug targets to identify active sites against plant diseases.

5.6 Chpater 4- Computational Modeling and Simulation for drug design

Plant viruses are infectious only in the presence of the viral coat protein; therefore, an understanding of coat protein's function is important for defining viral replication mechanisms [**63**]. Virus resistance has been conferred on plants by the expression of viral capsid or coat protein (CP) genes, replicase and movement protein genes, and viral antisense RNA [**64**]. Of these approaches, CP-mediated resistance has been most widely investigated and is near commercialization [**65**]. Capsid or coat proteins of many other plant viruses have been incorporated into the genomes of plants to provide protection [**66**]. This approach is based on the phenomenon of cross-protection, whereby a plant infected with a mild strain of virus is protected against a more severe strain of the same virus [**67**]. The phenotype of the resistant transgenic plants includes fewer centers of initial virus infection, a delay in symptom development, and low virus accumulation. Protoplasts from virus resistant transgenic plants are also resistant, suggesting that the protection is largely operational at the cellular level. Transgenic plants expressing viral coat protein are protected against infection by virus particles but are susceptible to viral RNA, indicating that the protection may primarily involve an inhibition of virus uncoating [**68**]. Proteins of viruses are necessary for its production in or on all food commodities. An exemption from the requirement of a tolerance is established for residues of the biological plant pesticide.

The new paradigm in vaccine design is emerging, following essential discoveries in immunology and development of new MHC Class binding peptides_prediction tools. MHC molecules are cell surface glycoproteins, which take active part in host immune reactions [**69**]. The involvement of MHC class in response to almost all antigens and the variable length of interacting peptides make the study of MHC Class 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 [**70-72**]. 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.

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.

Distance matrix based 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. 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.

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. 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.

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.

Predicted epitopes length and position are varying for different tools and results are not seen in one specific tool. Findings are the accurate results as the standard antigen, we used 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 antigenicity scales. These scales were designed to predict the locations of antigenic determinants. The Robson-Garnier and SOPMA method

predicted the secondary structure of coat protein. This methods show high antigenic regions present in beta sheet response than helical region of this peptide.

The physical and chemical nature of the proteins offers an approach to the analysis of function. For the study of physicochemical parameters, we consider 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. These scales are essentially a hydrophilic index, with apolar residues assigned negative values.

The protein sequence of plant viruses were analyzed and characterized to study the MHC class peptide binding, which allows potential drug targets to identify active sites against viral reactions. 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. 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.

The Support Vector Machine (SVM) based method is for prediction of promiscuous MHC class II binding peptides. 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.

For prediction of CTL epitopes we used, which are play very important role in subunit vaccine design. The neural network implementation has been achieved by using Stuttgart Neural Network Simulator, SNNS 4.2. These SVM and ANN based prediction methods were combined to establish the upper limit of sensitivity, specificity, accuracy. In consensus prediction, the accuracy of the prediction achieved. Antigenic epitopes of coat protein are important antigenic determinants against the viral attack.

Molecular biology techniques have enabled sequencing of number of proteins, but obtaining their three dimensional experimentally resolved structure is still not feasible for all proteins. Thus, alternative strategies are being applied to develop models of proteins. Among these, homology modeling is one of the methods being used most widely. Proteins are modeled using the experimentally resolved structure of the closest homologue, selected based on the percentage identity and similarity of the alignment. Modeling proteins using the closest homologue helps to unveil the structural features of such proteins and the nature of interactions with their ligand, thus enabling structure based drug design.

6. Data sources for the evaluation of this research work

Plant virus database provides a central source of information about 1837 viruses, viroids and satellites of plants, fungi and protozoa. Comprehensive taxonomic information, including brief descriptions of each family and genus, and classified lists of virus sequences are provided. The database also holds detailed, curated, information for all plant viruses, viroids and satellites of plants, fungi and protozoa that are complete. The evaluation of accuracy of prediction method is necessary to estimate the performance of a method.

Total 354 viruses' data were originally published in paper form by the Association of Applied Biologists (AAB) between 1970 and 1989, while additional descriptions have been added to a CD-ROM (also published by the AAB) since 1998. Virus descriptions are generated from database of the (International Committee of Taxonomy of Viruses) ICTV approved orders, families, subfamilies, genera and their type species, as well as many more species, strain and isolate descriptions.

Some new entries are taken from National Centre for Biotechnology Information (NCBI) Database. The descriptions can give from the indexes in the plant virus database menu until 1965 to 2008. New descriptions are being commissioned and will be added as they become available. The database is presented on the Visual Studio as a natural language output that is styled and connected to MS-Access database.

Regular searches are made at various resources for virus sequences that are new or that have been updated. Curated protein sequences are listed under the current species name from the appropriate genus, proteomics data are obtained from GenBank, other publicly available databases and websites, as well as from ongoing sequencing projects. The accession number is linked to the entire plant virus information file. Next to each accession number is a check box for selection of one or more sequences within the genus.

Evaluate lengths of ORFs fro given nucleotide sequences are the average distance between stop codons in "random" DNA is 64 / 3 ≈ 21, much smaller than the number of codons in an average protein (≈300). Fifth-order Hidden Markov Model algorithm would then take a given DNA sequence and search within each of the possible reading frames stop codons and its corresponding start codon. For each such potential ORF determine the length and evaluate that.

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.

Molecular phylogeny to retrace evolutionary relationships between sequences, which are taken from NCBI, EMBL, PIR, UniProt, and other publicly available databases, websites, as well as from ongoing sequencing projects. 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.

Automated neural-network based protein modeling server (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 the template file from sequence and used as a starting point for the homology modeling.

Automatic modeling of protein three-dimensional structure (Geno3D) is an automatic web server for protein molecular modeling. Starting with a query protein sequence, the server performs the homology modeling.

INSIGHT II is useful tool for structure based design and it is a powerful method for rapidly identifying new lead compounds when a receptor structure is available. We identified binding site, which locates protein binding site and uses sequence family information to identify characteristics that determine function. Also CHARMm is used to study the comprehensive force field and program for energy calculations, 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.

7. Summary

A proteomics research of plant diseases requires analysis of large sequences and is heavily dependent upon bioinformatics computation. Computational modeling should be performed according to strict standards, requiring careful data selection for protein model building, followed by adequate testing and validation through bioinformatics tools. The computational integration of such data is proving to be the most effective route to determine the protein function very quick and accurately. A database can provide easy access to previous results from the experiments and literature surveys, preventing the wasteful duplication of research. A well-designed database also supports both expert and machine guided searches for novel correlation in data. Integration of large datasets and information reveals from same resources in the life sciences is one of the most challenging goals facing biotechnology and pharmaceutical industries today.

8. References

[1] Jonathan Pevsner, Bioinformatics and Functional Genomics (John Wiley & Sons (Asia) Pte. Ltd, ISBN: 978-0-471-21004-7, 2004).

[2] Stephen Misener and Stephen A. Krawetz, Bioinformatics methods and protocols (Humana Press; 1 edition, ISBN-10: 0896037320, 1999).

[3] Christine Orengo, David Jones, Bioinformatics: genes, proteins and computers (Janet Thornton, Routledge, UK, ISBN: 978-1-85996-054-7, 2002).

[4] V.S. Gomase, S. Tagore, S.S. Changbhale and K.V. Kale, "Pharmacogenomics", *Current Drug Metabolism*, 9(3), 207-212, 2008. [Impact Factor- 5.76]-[PMID: 18336223]

[5] V.S. Gomase, K.V. Kale, S. Tagore and S.R. Hatture, "Proteomics: Technologies for Protein Analysis", *Current Drug Metabolism*, 9(3), 213-220, 2008. [Impact Factor- 5.76]-[PMID: 18336224]

[6] B.D. Harrison, "Usefulness and limitations of the species concept for plant viruses", *Intervirology*, 24(2), 71-78, 1985.

[7] A. Brunt, K. Crabtree, M. Dallwitz, A. Gibbs and L. Watson, Viruses of Plants: Descriptions and Lists from the VIDE Database (1484 pp. C.A.B. International, U.K., 1996).

[8] M.J. Adams, J.F. Antoniw, H. Barker, A.T. Jones, A.F. Murant, D. Robinson, Descriptions of Plant Viruses on CD-ROM, (Wellesbourne, Warwick, UK: Association of Applied Biologists; 1998).

[9] N. Kumar, B.S. Hendriks, K.A. Janes, D. de Graaf, D.A. Lauffenburger, "Applying computational modeling to drug discovery and development", *Drug Discovery Today*, 11(17–18), 806–811, 2006. [PMID: 16935748]

[10] D.M. Klinman, M. Takeno, M. Ichino, M. Gu, G. Yamshchikov, G. Mor, J. Conover, "DNA vaccines: safety and efficacy issues", *Springer Semin. Immunopathol*., 19(2):245-256, 1997.

[11] E. Hughes, H.J. Gilleland, "Ability of synthetic peptides representing epitopes of outer membrane protein F of Pseudomonas aeruginosa to afford protection against P. aeruginosa infection in a murine acute pneumonia model", *Vaccine*, 13(18), 1750-1753, 1995.

[12] J. Pellequer, E. Westhof, M. Van Regenmortel, "Predicting the location of continuous epitopes in proteins from their primary structure", *Methods Enzymol.*, 203, 176-201, 1991.

[13] M. Bhasin and G.P.S. Raghava, "Prediction of CTL epitopes using QM, SVM and ANN techniques", *Vaccine*, 22, 3195-3201, 2004.

[14] L.J. Stern and D.C. Wiley, "Antigen peptide binding by class I and class II histocompatibility proteins", *Structure*, 2, 245, 1994.

[15] P.A. Reche, J.P. Glutting and E.L. Reinherz, "Prediction of MHC Class I Binding Peptides Using Profile Motifs", *Human Immunology*, 63, 701-709, 2002.

[16] M. Kanehisa, S. Goto, "KEGG: kyoto encyclopedia of genes and genomes", *Nucleic Acids Res*., 28(1), 27-30, 2000.

[17] D. Fellmann, J. Pulokas, R.A. Milligan, B. Carragher and C.S. Potter, "A relational database for cryoEM: experience at one year and 50 000 images". *J. Struct. Biol*. 137, 273-282, 2002.

[18] M.J. Adams, and J.F. Antoniw, "DPVweb: a comprehensive database of plant and fungal virus genes and genomes", *Nucleic Acids Research*, 34(Database issue), D382-D385, 2006.

[19] J. Wang, J. Luo, Y. Li, H. Qu, G. Wu, X. Gu, "Construction of rice dwarf virus genome database", *Wei Sheng Wu Xue Bao*., 41(1), 43-48, 2001.

[20] M. Hirahata, T. Abe, N. Tanaka, Y. Kuwana, Y. Shigemoto, S. Miyazaki, Y. Suzuki, H. Sugawara, "Genome Information Broker for Viruses (GIB-V): database for comparative analysis of virus genomes", *Nucleic Acids Res*., 35(Database issue), D339-D342, 2007.

[21] Y. Huang, S.K. Lau, P.C. Woo, K.Y. Yuen, "CoVDB: a comprehensive database for comparative analysis of coronavirus genes and genomes", *Nucleic Acids Res*., 36(Database issue), D504-D511, 2008.

[22] M.J. Adams, J.F. Antoniw, H. Barker, A.T. Jones, A.F. Murant, D. Robinson, Descriptions of Plant Viruses on CD-ROM, (Wellesbourne, Warwick, UK: Association of Applied Biologists; 1998).

[23] R.K. Azad, M. Borodovsky, "Effects of choice of DNA sequence model structure on gene identification accuracy", *Bioinformatics*, 20(7), 993-1005, 2004.

[24] V. Brendel, P. Bucher, I. Nourbakhsh, B.E. Blaisdell, S. Karlin, "Methods and algorithms for statistical analysis of protein sequences", *Proceedings of the National Academy of Sciences USA 89,* 2002-2006, 1992.

[25] W. Kabsch, C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features", *Biopolymers*, 22(12), 2577-2637, 1983.

[26] G. Vriend, "WHAT IF: A molecular modeling and drug design program", *J. Mol. Graph*., 8(1), 52-56, 1990.

[27] B. Rost, G. Yachdav, J. Liu, "The PredictProtein server", *Nucleic Acids Res*., 32(Web Server issue), W321-W326, 2004.

[28] C.T. Zhang, R. Zhang, "A graphic approach to evaluate algorithms of secondary structure prediction", *J. Biomol. Struct. Dyn*., 17(5), 829-842, 2000.

[29] D.T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices", *J. Mol. Biol*., 292, 195-202, 1999.

[30] K.C. Parker, M.A. Bednarek and J.E. Coligan, "Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains", *J. Immunol.*, 152, 163, 1994.

[31] T. Sturniolo, E. Bono, J. Ding, L. Raddrizzani, O. Tuereci, U. Sahin, M. Braxenthaler, F. Gallazzi, M.P. Protti, F. Sinigaglia, J. Hammer, "Generation of tissue-specific and promiscuous HLA ligand database using DNA microarrays and virtual HLA class II matrices", *Nat. Biotechnol*., 17, 555-561, 1999.

[32] A.S. Kolaskar, P.C. Tongaonkar, "A semi-empirical method for prediction of antigenic determinants on protein antigens", *FEBS Lett*., 276(1-2), 172-174, 1990.

[33] S. Goldsmith-Fischman and B. Honig, "Structural genomics: Computational methods for structure analysis", *Protein Science*, 12, 1813–1821, 2003.

[34] C.B. Burge and S. Karlin, "Finding the genes in genomic DNA". *Curr. Opin. Struct. Biol*., 8, 346–354, 1998.

[35] S. Eddy, "Non-coding RNA genes". *Curr. Opin. Genet. & Dev*., 9, 695–699, 1999.

[36] D. Haussler, "Computational genefinding". *Bioinformatics: A Trends Guide*, 5,12–15, 1998.

[37] S. Rogic, A.K. Macksworth and F.B.F. Oulette, "Evaluation of gene-finding programs on mammalian sequences". *Genome Res*., 11, 817–832, 2001.

[38] L. Stein, "Genome annotation: from sequence to biology". *Nature Reviews Genet*., 2, 493–503, 2001.

[39] T. Strachan and A.P. Reid, "Identifying human disease genes". *In: Human Molecular Genetics 2, BIOS Scientific Publishers. Ltd, Oxford UK*, 351–375, 1999.

[40] I.M. Kapetanovic, "Computer-Aided Drug Discovery and Development (CADDD): in silico-chemico-biological approach". *Chem. Biol. Interact*., 171(2), 165–176, 2008.

[41] I. Dubchak, I. Muchnik, S.R. Holbrook, S.H. Kim, "Prediction of protein folding class using global description of amino acid sequence". P*roc. Natl. Acad. Sci. USA.*, 92, 8700–8704, 1995.

[42] Li Qingliang and Lai Luhua, "Prediction of potential drug targets based on simple sequence properties". *BMC Bioinformatics*, 8, 353, 2007.

[43] Shailza Singh, Balwant Kumar Malik and Durlabh Kumar Sharma, "Molecular drug targets and structure based drug design: A holistic approach". *Bioinformation*, 1(8), 314–320, 2006.

[44] M.R. Hilleman. "Vaccines in historic evolution and perspective: a narrative of vaccine discoveries". *Vaccine*, 18, 1436–1447, 2000.

[45] P.D. Minor, "Polio eradication, cessation of vaccination and re-emergence of disease". *Nature Reviews Microbiology*, 2, 473–482, 2004.

[46] R. Noad and P. Roy, "Virus-like particles as immunogens". *Trends in Microbiology*, 11, 438–444, 2003.

[47] L.R. Haaheim, *et al.*, "*Virus vaccines. Chapter 5 in A Practical Guide to Clinical Virology, 2nd edn, Wiley*", 2002.

[48] W. Jiskootm, *et al*., "*Vaccines. Chapter 12 in Pharmaceutical Biotechnology, editors Crommelin D. J. A. et al., 2nd edn, Taylor and Francis*", 2002.

[49] S. Moelbert, E. Emberly, and C. Tang, "Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins". *Protein Sci*., 13(3), 752 – 762, 2004.

[50] Laurence Lins, Annick Thomas and Robert Brasseur, "Analysis of accessible surface of residues in proteins". *Protein Science*, 12, 1406-1417, 2003.

[51] Frank Eisenhaber and Patrick Argos, "Hydrophobic regions on protein surfaces: definition based on hydration shell structure and a quick method for their computation". *Protein Engineering*, 9 (12), 1121-1133, 1996.

[52] D.S. DeLuca and R. Blasczyk, "Implementing the modular MHC model for predicting peptide binding". *Methods Mol. Biol*., 409, 261-271, 2007.

[53] P, Dönnes, "Support vector machine-based prediction of MHC-binding peptides". Methods Mol. Biol., 409, 273-282, 2007.

[54] W. Liu, J. Wan, X. Meng, D.R. Flower, T. Li , “In silico prediction of peptide-MHC binding affinity using SVRMHC”. *Methods Mol. Biol*., 409, 283-291, 2007.

[55] S. Lata, M. Bhasin, G.P. Raghava, “Application of machine learning techniques in predicting MHC binders”. *Methods Mol. Biol*., 409, 201-215, 2007.

[56] K. Yu, N. Petrovsky, C. Schönbach, J.Y. Koh, V. Brusic, “Methods for prediction of peptide binding to MHC molecules: a comparative study”. *Mol. Med*., 8(3), 137-148, 2002.

[57] P. Donnes and A. Elofsson, “Prediction of MHC class I binding peptides, using SVMHC”. *BMC Bioinformatics*, 3, 25, 2002.

[58] A. Shrake and J.A. Rupley, “Environment and exposure to solvent of protein atoms. Lysozyme and insulin”. *J. Mol. Biol*., 79(2), 351–371, 1973.

[59] B. Lee and F.M. Richards, “The interpretation of protein structures: estimation of static accessibility”. *J. Mol. Biol*., 55(3), 379–400, 1971.

[60] D.A. Case, T.E. Cheatham, III, T. Darden, H. Gohlke, R. Luo, K.M. Merz, Jr., A. Onufriev, C. Simmerling, B. Wang and R. Woods, “The Amber biomolecular simulation programs”. *J. Computat. Chem*., 26, 1668-1688, 2005.

[61] J.W. Ponder and D.A. Case, “Force fields for protein simulations”. *Adv. Prot. Chem*., 66, 27-85, 2003.

[62] T.E. Cheatham, III and M.A. Young, “Molecular dynamics simulation of nucleic acids: Successes, limitations and promise”. *Biopolymers*, 56, 232-256, 2001.

[63] M. Laura, Siana M. Laforest, and Lee Gehrke, “Coat Protein Activation of Alfalfa Mosaic Virus Replication Is Concentration Dependent”. *Journal of Virology* 79, 9, 5752-5761, 2005.

[64] T.M.A. Wilson, “Strategies to protect crop plants against viruses: pathogen-derived resistance blossoms”. *Proc. Natl. Acad. Sci. USA*, 90, 3134-3141, 1993.

[65] M. Bendahmane, I. Chen, S. Asurmendi, A.A. Bazzini, J. Szecsi, R.N. Beachy, “Coat protein-mediated resistance to TMV infection of Nicotiana tabacum involves multiple modes of interference by coat protein”. *Virology,* 366(1), 107-116, 2007.

[66] L.S. Loesch-Fries, D. Merlo, T. Zinnen, L. Burhop, K. Hill, K. Krahn, N. Jarvis, S. Nelson, E. Halk, "Expression of alfalfa mosaic virus RNA 4 in transgenic plants confers virus resistance". *EMBO J*., 6(7), 1845-1851, 1987.

[67] J.L. Sherwood and R.W. Fulton, "The specific involvements of coat protein in tobacco mosaic virus cross protection". *Virology*, 119, 150-158, 1982.

[68] Vidadi Yusibov and L. Sue Loesch-Fries, "High-affinity RNA-binding domains of alfalfa mosaic virus coat protein are not required for coat protein-mediated resistance (cross-protection/engineered resistance)". *Proc. Natl. Acad. Sci. USA*, 92, 8980-8984, 1995.

[69] M. Bhasin and G.P. Raghava, "Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences". *Nucleic Acids Res*., 33, W202-7, 2005.

[70] S. Buus, S.L. Lauemoller, P. Worning, C. Kesmir, T. Frimurer, S. Corbet, A. Fomsgaard, J. Hilden, A. Holm, S. Brunak, "Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach". *Tissue Antigens,* 62, 378-384, 2003.

[71] M. Nielsen, C. Lundegaard, P. Worning, S.L. Lauemoller, K. Lamberth, S. Buus, S. Brunak, O. Lund, "Reliable prediction of T-cell epitopes using neural networks with novel sequence representations". *Protein Sci*., 12, 1007-1017, 2003.

[72] M. Nielsen, C. Lundegaard, P. Worning, C.S. Hvid, K. Lamberth, S. Buus, S. Brunak, O. Lund, "If you specifically use the weight matrix derived predictions, please also cite: Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach". *Bioinformatics*, 20(9), 1388-1397, 2004.