IMMUNO-INFORMATICS

The human immune system is very complex and operates at multiple levels, molecules, cells, organs, and organisms. Each individual has a unique immune system and will respond differently to immune challenges. It has a combination of biological structures and processes within an organism to protect it against disease. The earliest literary reference to immunology goes back to 430b.c., courtesy Thucydides. In 1798, Edward Jenner found some milkmaids immune to smallpox because earlier they contacted cowpox (a mild disease). The next major advancement in immunology came with the induction of immunity to cholera by Louis Pasteur. After applying weakened pathogen to animals, he administered a dose of vaccine to a rabid dog-bitten boy who later survived. But Pasteur could not explain its mechanism. In 1890, experiments of Emil Von Behring and Shibasaburo Kitasato led to the understanding of the mechanism of immunity. Their experiments described that antibodies present in the serum provided protection against pathogens. According to the traditional dogma of immunology, vertebrates have both innate and adaptive immunology. Innate immune system acts more rapidly and is older and more evolutionarily conserved in comparison with adaptive immune system. It provides the backbone on which adaptive immune system was able to evolve. Innate immune system is less specific and works as a first line of defence. It comprises four types of defensive barriers, viz., anatomic (e.g., skin and mucous membranes), physiologic (e.g., temperature, low pH), phagocytic (e.g., blood monocytes, neutrophils, tissue macrophages), and inflammatory (e.g., serum proteins).

Adaptive immune responses in vertebrates are generated within 5 or 6 days after the initial exposure to the pathogen. It is coordinated by a network of highly specialized cells that communicate through cell surface molecular interactions and a complex set of intercellular communication molecules known as cytokines and chemokines. Later exposure to the same pathogen induces a heightened and more specific response because it retains memory. Adaptive immune system has two parts: the cellular immune response of T cells and humoral response of B cells. An antigen has a specific small part, known as epitope that is recognized by the corresponding receptor present on B or T cells. B cell epitopes can be linear and discontinuous amino acids. T cell epitopes are short linear peptides. Most of the T cells can be in either of the two subsets, distinguished by the presence of one or the other of the two glycoproteins on their surface, designated as CD8 or CD4. CD4 T cells function as T helper (Th) cells that recognize peptides displayed by MHC class II molecules. On the other hand, CD8 functions as Tc (cytotoxic T) cells which recognize peptides displayed by MHC class I molecules. The complexity of the immune system arises from its hierarchical and combinatorial properties. Thus, huge amount of data related to immune systems is being generated. Immunologic research needs to deal with this complexity. Immunologists have been using high throughput experimental techniques for quite a long time, which have generated a vast amount of functional, clinical, and epidemiological data. Therefore, the development of new computational approaches to store and analyze these data is needed. This gives rise to the field called immuno-informatics. Immunogenomics, immunoproteomics, epitope prediction, and in silico vaccination are different areas of computational immunological research. Recently, systems biology approaches are being applied to

investigate the properties of dynamic behavior of an immune system network.

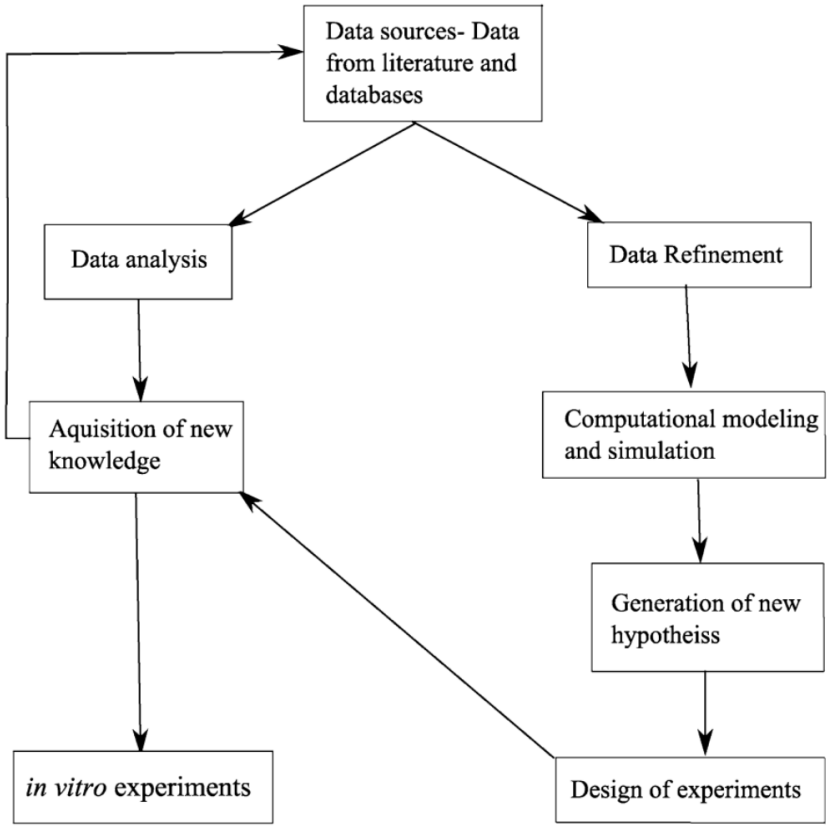

Fig. 1 A possible work flow in immunomics

It includes the study and design of algorithms for mapping potential B and T cell epitopes. These also can lead to exploring the potential binding sites for the development of new vaccines. This methodology is termed as “reverse vaccinology”. It is quite advantageous because conventional methods need to cultivate pathogen and then to extract its antigenic proteins. All the genes and proteins taking part in immune responses are referred to as “immunome,” and it excludes genes and proteins that are expressed in cell types other than in immune cells. All immune reactions due to interaction between host and antigenic peptides are referred to as “immunome reactions,” and their study is called as “immunomics”. Like genomics and proteomics, immunomics is a new discipline, which uses high-throughput techniques to understand immune system mechanism. Figure 1 shows work flow in immunomics. This chapter describes various available information regarding classical immunology, different immunomic databases, B and T cell epitope prediction tools and software, and applications of immuno-informatics.

Data Sources

Data sources include availability of data from lab experiments through scientific literature, molecular databases, tools and web servers, and clinical records. In this part of the lecture, we focus on various immune system-related data types and databases. The section starts with some experimental techniques and results.

Data from Lab Experiments

Immunological experimental and high-throughput molecular biology techniques help in finding the structure and function of immune genes and their products and thereby accumulating a vast amount of experimental data. Experiments involve many immunological techniques to understand the mechanism of an immune system and its responses to various infections, diseases, and drugs, viz., affinity chromatography, flow cytometry, radio-immunoassay (RIA), enzyme-linked immunosorbent assay (ELISA), competitive inhibition assay, and Coombs test. Here, we present some experimental findings that help to identify B and T cell epitopes and to study immune responses. The ability to identify epitopes in the immune response has important implications in diagnosis of diseases. Thus, epitopes for B and T cells need to be identified and mapped. In this context, Wanga et al. mapped B cell epitope present on non-structural protein (NS1), viz., NS1-18 and NS1-19, in Japanese encephalitis virus. For epitope mapping, a series of 51 partially overlapping fragments covering the entire NS1 protein were expressed with a glutathione *S*-transferase (GST) tag and then screened by a monoclonal antibody (mAb). Purification techniques like affinity chromatography are used to purify MHC–peptide from membrane MHC molecules, which can be analyzed by capillary high-pressure liquid chromatography electrospray ionization-tandem mass spectrometry. They can be further used to find new tumour-associated antigens (TAA). One such approach to find TAA is based on transfection of expression library made from cDNA into cells expressing the desired MHC haplotypes. The clones are selected on the basis of their ability to provoke immune response in T cells of the individuals with the same MHC type.

Exploring the Microarray Technology for Immunomics

“Immunomic microarray” is a microarray technique based on the principle of binding and measurement of target biological specimens to complementary probes. It helps in selecting proteins that cause autoimmunity from genomic sequences. It is being applied to autoimmune disease diagnosis and treatment, allergy prediction, T and B cell epitope mapping, and vaccination to name a few. It includes dissociable antibody microarray, serum microarray, and serological analysis of cDNA expression library (SEREX). An antibody microarray is used to measure concentration of antigen for a specific antibody probe and thereby consists of antibody probes and antigen targets. On the contrast, peptide microarray uses antigen peptides as fixed probes and serum antibodies as targets. The recent technology is peptide–MHC microarray or artificial antigen-presenting chip. In this technique, recombinant peptide–MHC complexes and co-stimulatory molecules are immobilized on a surface, and population of T cells is incubated with the microarray. The T cell spots act as artificial antigen-presenting cells containing a defined MHC-restricted peptide. The advantage of using peptide–MHC is that it can map MHC-restricted T cell epitope. The immunomic and genomic microarray data have some similarities, yet both of them also differ in several ways; for example, both of them have different designs. One can measure two or more signals simultaneously determined by a single feature, i.e., epitope in immunomic microarray. DNA microarrays measure one response value for each gene per sample; that is, mRNA

concentration produced by the gene but a single epitope can generate different response values corresponding to different epitopes in peptide–MHC chips. In case of B cell epitope, it can be recognized by different isotypes of immunoglobulins, so here, one can measure both intensity and quality of antibody response.

Immunomic Databases

The property of an antigen to bind specifically complementary antibodies is known as the antigen's antigenicity. Likewise, the ability of an antigen to induce an immune response is called its immunogenicity. Immunomic databases include epitope information-related databases, analysis tools, and prediction algorithms, which are crucial for basic immunological studies, diagnosis, and treatment of various diseases and in vaccine research. InnateDB (http://www.innatedb.ca) has been created to understand complete network of pathways and interactions of innate immune system responses. It has ~18,000 annotated molecular interactions of relevance to innate immunity and >1,200 genes, involved in innate immunity according to the recent update till February 16, 2012. It has a newer version, called Cerebral, which is a Java plug-in for the cytoscape biomolecular interaction viewer version 2.8.2 for automatically generating layouts of biological pathways. Table 1 lists some of the databases that deal with information related to B cell epitopes, T cell epitopes, allergy prediction, and evolution of immune system genes and proteins.

B Cell Epitope Databases

A brief detail on B cell epitope databases is provided here. Mapping B cell epitopes plays an important role in vaccine design, immunodiagnostic tests, and antibody production. It has been found that 90 % of B cell epitopes are conformational or discontinuous; however, they may comprise linear amino acid chain of peptides, which is brought closure in 3D space. Bcipep (http://www.imtech.res.in/raghava/bcipep) gives comprehensive information about experimentally verified B cell epitopes and tools for mapping these epitopes on an antigen sequence. Conformational epitope database (CED) has a collection of B cell epitopes from the literature, conformational epitopes defined by methods, like X-ray diffraction, NMR, scanning mutagenesis, overlapping peptides, and phage display.

Epitome (http://www.rostlab.org/services/epitome/) contains all known antigen–antibody complex structures. A semiautomated tool has also been developed which identifies the antigenic interactions within the known antigen–antibody complex structures. They compiled these interactions into Epitome. None of the other databases till now explicitly can locate the complementary determining regions (CDRs) or identify the antigenic residues semiautomatically. Epitome update follows update of SCOP; that is, Epitome is updated twice a year as soon as SCOP gets updated. The difference between Epitome and CED lies in the source of collection of B cell epitopes. Epitome collects B cell epitopes only from PDB structures and includes CDR information. In contrast, CED takes data from the literature and from abovementioned methods. As their sources are different, one can use the

complementary information.

Table 1
Databases on B cell epitopes, T cell epitopes, allergen, and molecular evolution of immune system components

Databases	Names	URLs
B cell epitopes	CED Bcipep Epiotme IEDB IMGT®	http://www.immunet.cn/ced/log.html http://www.imtech.res.in/raghava/bcipep http://www.rostlab.org/services/epitome/ http://www.immuneepitope.org/ http://www.imgt.org
T cell epitopes	Syfpeithi IEDB IMGT®	http://www.syfpeithi.de http://www.immuneepitope.org/ http://www.imgt.org
Allergen	Database of IUIS SDAP	http://www.allergen.org http://www.fermi.utmb.edu/SDAP/
Information related to molecular evolution of immune system components	ImmTree Immunome database ImmunomeBase Immunome Knowledge Base	http://www.bioinf.uta.fi/ImmTree http://www.bioinf.uta.fi/Immunome/ http://www.bioinf.uta.fi/ImmunomeBase http://www.bioinf.uta.fi/IKB/

T Cell Epitope Databases

A brief detail on T cell epitope databases is provided here. A detailed description can be found in later chapters. A functional T cell response requires MHC–peptide binding and a proper interaction of the MHC–peptide ligand with a specific T cell receptor. We need well-characterized data to model the process of binding of peptides to TAP and MHCs which function as T cell epitopes. Some recent investigations include finding and mapping of potential epitopes. Epitope mapping leads to designing effective vaccines.

Syfpeithi database (http://www.syfpeithi.de) has information on MHC class I and II anchor motifs and binding specificity. It calculates a score based on the following rules — calculated score values differentiate among anchor, auxiliary anchor, or preferred residues. IEDB has more than 88382 peptidic epitopes and can be found at http://www.immuneepitope.org/ and ontology-related information (http://ontology.iedb.org/) which has been specifically designed to capture intrinsic, chemical, and biochemical information on immune epitopes and their interactions with molecules of the host immune system. A beta version of IEDB (Immune Epitope Database and Analysis Resource Database) (http://www.immuneepitope.org/), sponsored by the National Institute for Allergy and Infectious Diseases (http://www.niaid.nih.gov) (NIAID), has different tools to find B and T cell epitopes. It had 88382 peptidic epitopes till February 2012. FRED deals with the methods for data processing and to compare the performance of the prediction methods considering experimental values. IMGT® (the international ImMunoGeneTics information system®) (http://www.imgt.org) has a good collection of IG, TR, MHC, and related proteins of the immune system of human

and other vertebrates. It has five databases and 15 interactive online tools for sequence, genome, and 3D structure analysis. The IMGT/HLA Database (http://www.ebi.ac.uk/imgt/hla) provides a specialist database that has 5,518 HLA class I alleles and 1,612 HLA class II alleles. It is a part of the international ImMunoGeneTics project (IMGT).

Allergy Prediction Databases

Allergy is a steadily increasing health problem for all age groups caused by allergens. Allergens are proteins or glycoproteins recognized by IgE that is produced by the immune system in allergic individuals. Online allergen databases and allergy prediction tools are being used to find cross-reactivity between known allergens. Localization of B and T cells in the allergen may not coincide. The differences between both kinds of epitopes present in an antigen are as follows: T cell epitopes are only linear (as mentioned earlier) and distributed throughout the primary structure of the allergen, whereas B cell epitopes can be either linear or conformational, recognized by IgE antibodies, and are located on the surface of the molecule accessible to antibodies. Moreover, in the case of B cell epitopes, predicting allergenicity in a molecule based on known conformational epitopes is a difficult task.

Immunomic Tools and Algorithms

The property of an antigen to bind specifically complementary antibodies is known as the antigen's antigenicity; likewise, the ability of an antigen to induce an immune response is called its immunogenicity. The main objective of epitope prediction is to design a molecule that can replace an antigen in the process of either antibody production or antibody detection. Such a molecule can be synthesized or, in case of a protein, its gene can be cloned into an expression vector. Designed molecules are inexpensive and noninfectious in contrast to viruses or bacteria. Epitopes are important for understanding the disease mechanism, host–pathogen interaction analyses, antimicrobial target discovery, and vaccine design. Traditionally, determination of binding affinity of MHC molecules and antigenic peptides predicts epitopes. The experimental techniques are found to be difficult and time consuming. Due to this reason, several in silico methodologies are being developed and used to identify epitopes. Here, we throw some light on available immunology-related tools and algorithms. These techniques include matrix-driven methods, finding structural binding motifs, quantitative structure–activity relationship (QSAR) analysis, homology modeling, protein threading, docking techniques, and design of several machine-learning algorithms and tools. Table 2 lists some of the tools that deal with B and T cell epitope prediction, allergy

prediction, and in silico vaccination.

Table 2
Web servers and tools for prediction of B and T cell epitopes, allergens, and in silico vaccination

Web servers and tools	Names	URLs
B cell epitope prediction	ABCpred	http://www.imtech.res.in/raghava/abcpred
	COBEpro	http://www.scartch.proteomics.uci.edu
	Bepipred	http://www.cbs.dtu.dk/services/BepiPred
	IMGT®	http://www.imgt.org
	Bcepred	http://www.imtech.res.in/raghava/bcepred/
	DiscoTope	http://www.cbs.dtu.dk/services/DiscoTope/
	CEP	http://www.115.111.37.205/cgi-bin/cep.pl
	AgAbDb	http://www.115.111.37.206:8080/agabdb2/home.jsp
	MIMOP	Request from franck.molina@cpbs.univ-montp1.fr
	MIMOX	http://www.immunet.cn/mimox/
	Pepitope	http://www.pepitope.tau.ac.il/
	3DEX	http://www.schreiber-abc.com/3dex/
	IEDB	http://www.immuneepitope.org
T cell epitope prediction	MMBPred	http://www.imtech.res.in/raghava/mmbpred/
	NetCTL	http://www.cbs.dtu.dk/services/NetCTL/
	NetMHC 3.0	http://www.cbs.dtu.dk/services/NetMHC/
	TAPPred	http://www.imtech.res.in/raghava/tappred/
	Pcleavage	http://www.imtech.res.in/raghava/pcleavage/
	ElliPro	http://www.tools.immuneepitope.org/tools/ElliPro
	MHCPred	http://www.ddg-pharmfac.net/mhcpred/MHCPred/
	Propred	http://www.imtech.res.in/raghava/propred1/
	EpiToolKit	http://www.epitoolkit.org
	Syfpeithi	http://www.syfpeithi.de
	IMGT®	http://www.imgt.org
	IEDB	http://www.immuneepitope.org/
	EpiJen v 1.0	http://www.ddg-harmfac.net/epijen/EpiJen/EpiJen.htm
Allergy prediction	AlgPred	http://www.imtech.res.in/raghava/algpred
	Allermatch	http://www.allermatch.org
	APPEL	http://www.jing.cz3.nus.edu.sg/cgi-bin/APPEL
	EVALLER	http://www.slv.se/en-gb/Group1/Food-Safety/e-Testing-of-protein-allergenicity/
In silico vaccination	VaxiJen	http://www.ddg-pharmfac.net/vaxijen/
	DyNAVacs	http://www.miracle.igib.res.in/dynavac/
	NERVE	http://www.bio.unipd.it/molbinfo
	VIOLIN	http://www.violinet.org
	Vaxign	http://www.violinet.org/vaxign/

B Cell Epitope Prediction

Experimental determination of B cell epitopes is time consuming and expensive; there is a need for computational methods for reliable identification of putative B cell epitopes from antigenic sequences. B cell epitopes are antigenic determinants on the surface of pathogens that interact with B cell receptors (BCRs). BCRbinding site is hydrophobic, having six hypervariable loops of variable length and amino acid composition. B cell epitopes

are classified as continuous/linear/sequential and discontinuous conformational. Linear epitopes are short peptides that correspond to a contiguous amino acid sequence fragment of a protein. However, most epitopes are discontinuous, where distant residues are brought into spatial proximity by protein folding within the folded 3D protein structure. Experiments are mostly based on linear epitopes. There are both sequence-based and structure-based prediction tools, but prediction tools are limited for discontinuous B cell epitopes.

Prediction of Methodology for Continuous B Cell Epitopes

Methodologies for prediction of continuous B cell epitopes involve sequence-based methods, amino acid propensity scale-based methods, and machine-learning methods.

Sequence-Based Methods

Sequence-based methods generally look for the epitope surface that must be accessible for antibody binding. These methods are limited to the prediction of continuous epitopes. Sequence-based methods have been tested on prediction of two protective epitopes known in influenza A virus hemagglutinin HA1. The first continuous epitope is the 91–108 epitope (SKAFSNCYPYDVPDYASL), which is a protective epitope in rabbit able to elicit antibodies neutralizing infectivity of influenza viruses. The second continuous epitope is the 127–133 epitope (WTGVTQN) protective against the influenza strain A/Achi/2/68 (H3N2) in mouse.

Amino Acid Propensity Scale-Based Methods

Parameters such as hydrophilicity, flexibility, accessibility, turns, exposed surface, polarity, and antigenic propensity of polypeptide chains have been correlated with the location of continuous epitopes. Thus, the classical methods of identifying potential linear B cell epitopes from antigenic sequences typically rely on the use of amino acid propensity scales. Amino acid scale-based methods apply amino acid scales to compute the scores of a residue *i* in a given protein sequence. The *i*-(*n*-1)/2 neighboring residues on each side of residue *i* are used to compute the score for residue *i* in a window of size *n*. The final score for residue *i* is the average of the scale values for *n* amino acids in the window. Pellequer compared several propensity scale methods using a dataset of 14 epitope-annotated proteins.

Machine-Learning Methods

Machine-learning algorithms and tools are being used to retrieve characteristics of an epitope. Here we describe some of these approaches in brief. Saha and Raghava used feed-forward and recurrent neural networks to predict continuous B cell epitopes in ABCpred (http://www.imtech.res.in/raghava/abcpred). COBEpro is a two-step system for prediction of continuous B cell epitopes. In the first step, COBEpro assigns a fragment epitopic propensity score to protein sequence fragment using SVM. In the second step, it calculates an epitopic propensity score for each residue based on the SVM scores of the peptide fragment in the antigenic sequence. For Bepipred, (http://www.cbs.dtu.dk/services/BepiPred), three datasets

of linear B cell epitopes were constructed, viz., annotated proteins from literature, AntiJen database (http://www.ddg-pharmfac.net/antijen/AntiJen/antijenhomepage.htm), and Los Alamos HIV database (http://www.hiv.lanl.gov). They tested a number of propensity scale methods on Pellequer dataset and found the best scale by Levitt. Then, they used HMM to predict the location of linear B cell epitopes and tested HMMs on Pellequer dataset to find optimal parameters. HMM was combined with one set of the two best propensity scale methods, i.e., Parker and Levitt, to get the more accurate predictions. Currently, ~60–66 % of accuracy has been found for continuous epitope prediction, applying combinations of either amino acid scales or machine-learning techniques. The higher accuracy could possibly be achieved by improving the quality of existing B cell epitope datasets.

Prediction Methodology for Discontinuous B Cell Epitopes

The characterization and prediction of B cell epitopes are mainly conformational dependent based on the knowledge of the protein three-dimensional structure; thus the task of prediction is more difficult compared to that of T cell epitopes. Changes in protein folding may lead to changes in the number of epitopes. The most accurate way to identify B cell epitope is through X-raycrystallography. Here we describe some of the prediction methods for conformational B cell epitopes in brief. Anderson et al. presented a method called DiscoTope (http://www.cbs.dtu.dk/services/DiscoTope/), which is a combination of amino acid statistics, spatial information, and surface exposure. It detects 15.5 % of residues

located in discontinuous epitopes with a specificity of 95 %. It is said to be the first method developed for prediction of discontinuous B cell epitope with better performance than methods based only on sequence data. PEPITO uses a weighted linear combination of amino acid propensity scores and half-sphere exposure values which encode side chain orientation and solvent accessibility of amino acid residues for the prediction of conformational epitopes. Authors have also reported its improvement in performance over DiscoTope method. Bublil et al. developed Mapitope for conformational B cell epitope mapping. The hypothesis behind Mapitope is that the simplest meaningful fragment of an epitope is an amino acid pair (AAP) of residues that lie within the epitope, which are the results of folding. A set of affinity-isolated peptides was obtained by screening the phage display peptide libraries with the antibody of interest. This set was given as algorithm input, and 1–3 epitope candidates on the surface of the atomic structure of the antigens were obtained as output. A computational method has been presented by Sollner et al. to automatically select and rank peptides for the stimulation of otherwise functionally altered antibodies. They investigated the integration of B cell epitope prediction with the variability of antigen and the conservation of patterns for posttranslational modification (PTM) prediction. By their observation, they found high antigenicity, low variability, and low likelihood of PTM for the identification of biorelevant sites.

Mimotope-Based Methodology

Phage display library is widely used for finding protein–protein interactions (specially in antibody–antigen interactions), protein function identification, and development of new drugs and vaccines. Pizzi et al. have proposed an approach for mapping B cell epitopes, in which a

phage display library of random peptides is scanned against a desired antibody to obtain mimotopes that bind to the antibody with high affinity. It is assumed that this panel of mimotopes mimics the physicochemical properties and spatial organization of the genuine epitopes. Mimotopes and antigens are both recognized by the same antibody paratope. Mimotopes are said to be the imitated part of the epitope. It is possible that mimotope may have some valuable information about epitope. However, homology may not exist between the mimotope and the epitope of the native antigen. This mimicry exists due to similarities in physiochemical properties and spatial organization.

Hybrid (Ensemble) Prediction Method

Ensemble methods combine the predictions of several predictors and often outperform individual predictors in many biomolecular sequence and structure classification studies. Several strategies for combining a set of predictors, *S*, into a single consensus or meta-predictor exist:

(1) majority voting,

(2) weighted linear combination, and

(3) meta-learning.

A large number of nearest neighbour - and decision tree-based classifiers are trained using different sets of training data features for developing an ensemble of linear B cell epitope classifiers.

T Cell Epitope Prediction

The current challenge in immunological prediction software is to predict interacting molecules to a high degree of accuracy. The most popular methods currently available are based on binding affinity predictions for a range of MHC molecules. It is necessary to bind antigenic peptides with MHC so that cytotoxic T cells can recognize them. Thus, identification of MHC-binding peptides is a central part of any algorithm which predicts T cell epitopes. There exist several methodologies for prediction of MHC-binding peptides, which are based on the idea of quantitative matrices, hidden Markov model (HMM), artificial neural networks (ANNs), support vector machine (SVM), and structure of the peptides.

Matrix-Driven Methods

Huang and Dai first investigated a new encoding scheme of peptides based on BLOSUM matrix with the amino acid indicator vectors for direct prediction of T cell epitopes. It replaced each nonzero entry in the amino acid indicator vector by the corresponding value appeared in the diagonal entries in BLOSUM matrix. MMBPred (http://www.imtech.res.in/raghava/mmbpred/) server predicts the mutated promiscuous and high affinity MHC-binding peptide. It uses the matrix data in a linear prediction model and ignores peptide conformation. The prediction is based on the quantitative matrices of 47 MHC alleles.

Hidden Markov Model-Based Method

Transfer-associated protein (TAP) is an important component of the MHC I antigen processing and presentation pathway. A TAP transporter can translocate peptides of 8–40 amino acids into endoplasmic reticulum (ER). Zhang et al. developed PREDTAP for the prediction of peptide binding to hTAP. They used a three-layer back propagation network with the sigmoid activation function. The inputs were the binary strings, representing nonamer peptide. Secondly, they used second-order HMM. The results are both sensitive and specific.

Artifi cial Neural Network-Based Method

ANNs can identify each amino acid residue and interactions between adjacent ones in a potential epitope. An ANN for a particular MHC molecule is trained to recognize associated input sequence and output., the binding affinity for that sequence with the MHC molecule. Trained ANN can predict the binding affinity of novel peptide sequences. Neilson et al. described an improved neural network model to predict T cell class I epitopes.

They combined a sparse encoding, BLOSUM encoding, and input derived from HMM. The dataset consists of 528 nine-mer amino acid peptides for which the binding affinity to the HLA I molecule A*0204 has been measured in a method described by Buus et al. NetCTL server (http://www.cbs.dtu.dk/services/NetCTL/) has method to integrate the prediction of peptide MHC class I binding, proteasomal C terminal cleavage, and TAP transport efficiency. NetMHC server 3.0 (http://www.cbs.dtu.dk/ services/NetMHC/) uses ANN and weight matrices. It has been trained on data from 55 MHC peptides (43 human and 12 nonhuman) and position-specific scoring matrices (PSSMs) for additional 67 HLA alleles. Prediction of MHC class II binding peptides is found to be difficult due to the reasons including variable length of reported binding peptides, undetermined core region for each peptide, and number of amino acids as primary anchor. Brusic et al. developed PERUN, a hybrid method for the prediction of MHC class II binding peptide. It uses available experimental data and expert knowledge of binding motifs, evolutionary algorithms, and ANNs. They used PlaNet package version 5.6 to design and train a three-layered fully connected feed-forward ANN. The whole process of MHC class I ligands' degradation and presentation has been modeled in EpiJen (http://www.ddg-pharmfac.net/epijen/EpiJen/EpiJen.htm) in an integrative approach. It is a multi-step algorithm for T cell epitope prediction, based on quantitative matrices, which belongs to the next generation of in silico T cell epitope identification methods.

Other Machine Learning Methods

Ant colony search systems (ACSs) have been found useful for solving combinatorial optimization problems and can be applied to the identification of a multiple alignment of a set of peptides. Basically, ACSs attempt to find an optimal alignment for a given set of peptides based on the search strategy. For TAPPred (http://www.imtech.res.in/raghava/tappred/), nine features of amino acids have been analyzed to find the correlation between binding affinity and physiochemical properties. An SVM-based method to predict TAP binding affinity of peptides has been developed and found cascade SVM to be more reliable. Cascade SVM

has two layers of SVMs, and its performance is better than the other available algorithms. Nanni demonstrated the use of SVM and support vector (SV) data description to predict T cell epitope. It is experimentally established that the immunoproteasome is involved in the generation of the MHC class I ligand. For this purpose, Pcleavage (http://www.imtech.res.in/raghava/pcleavage/) has been developed to predict both kinds of cleavage sites in antigenic proteins. It uses SVM, Parallel Exemplar based Learning (PEBLS), and Waikato Environment for Knowledge Analysis (Weka).

Structure-Based Prediction

Accurate identification of peptides that bind to specific MHC molecules is important for understanding the underlying mechanism of immune recognition, for developing effective peptide-based vaccines, and for immunotherapies for allergy and autoimmunity. Current methods are mostly based on peptide binding affinity to MHC for predicting T cell epitope. 3D QSAR technology CoMSIA has been applied to the problem of peptide–MHC binding. It uses the interaction potential around aligned sets of 3D peptide structures to describe binding. TEPITOPE is used to predict promiscuous and allele-specific HLA II-restricted T cell epitope in silico. TEPITOPE's user interface has a display and comparison of pocket profiles, and it finds similar HLA II differing in their binding capacity for a given peptide sequence. It can be applied to only 51 out of over 700 known HLA-DR molecules.

Molecular Dynamics-Based Prediction

Molecular dynamics (MD) describes single and collective motion of atoms within a molecular system and provides a means by which one can measure theoretically that cannot be measured experimentally. It is particularly suitable for the simulation and analysis of the otherwise inaccessible details of MHC–peptide interaction and of the immune synapse. Zhang et al. were among the first who uses MD as a tool to explore peptide–MHC binding. They focused on docking using MD as well as on calculating free energies. Free energy calculations of the wild-type and the variant human T cell lymphotropic virus type 1 Tax peptide (LLFGYPVYV—wild Tax and LLFGYAVYV—mutant Tax) presented by the MHC to the TCR have been performed using large-scale massively parallel molecular dynamics simulations.

Allergy Informatics

Allergy is caused by adverse immunological reaction, and the causative agents are known as allergens that are otherwise not harmful in nature. An allergen cross-links immunoglobulin E (IgE) antibody on mast cells or basophils and releases inflammatory mediators that cause allergy symptoms. Biotechnology- and genetic engineering-derived food contains some foreign proteins, which can be allergic to many human beings. Evaluation of the potential allergenicity of food derived from biotechnology and genetic engineering is a current food safety assessment. Allergen sequence databases are essential tools for safety assessments of bioengineered foods. They can analyze the structural and physiochemical properties of food allergen proteins. Current efforts in allergy informatics are primarily focused on prediction of T and B cell epitopes and assessment of allergenicity. Allergy occurs by both extrinsic and intrinsic factors. Type I hypersensitive reaction is induced by certain allergens that elicit

IgE antibodies. Use of genetically modified food and therapeutics makes allergenic protein prediction necessary. According to the proposed guidelines of World Health Organization (WHO) and Food and Agriculture Organization (FAO) in 2001, a protein that has at least six same contiguous amino acids or a window of 80 amino acids when compared with known allergens is considered as allergen. It has already been established that allergens do not share common structural characteristics. Thus, allergen databases are being used as reference for finding the sequence similarity in allergenicity evaluation. It is said that a protein is considered as an allergen if it has a region or peptides identical to a known IgE epitope. Allergen prediction method proposed by Kong et al. Is based on the determination of a combination of two allergen motifs in a given protein sequence. They took 575 proteins for allergen dataset and 700 sequences for non-allergen test set from the given reference. They developed a database which has all possible combinations of two motifs from the set of allergenic motifs by using motif length of 35 amino acids and motif number of 500. Zorzet et al. introduced a computational approach for classifying the amino acid sequences in allergens and nonallergens. They identified pre-processed 91 food allergens from various specialized public repositories of food allergy and SWALL database (SWISSPROT and TrEMBL). AlgPred (http://www.imtech.res.in/raghava/algpred) uses SVM and a similarity-based approach for analysis and scanned all 183 IgE epitopes against all proteins of the dataset. The server allows using a hybrid option to predict allergen using combined approach (SVMc, IgE epitope, ARPs BLAST, and MAST). Stadler et al. used MEME motif discovery tool to identify the most relevant motif present in allergen sequence. If the query finds an allergen motif or scores better than an E-value of 10-8 in the pairwise sequence alignment step, it is considered as the allergenic sequence. Then, these are compared with the FAO/WHO guidelines by performing allergenicity prediction for the sequence in SWISSPROT, and a synthetic test database ALLERMATCH (http://www.allermatch.org) is a web tool that uses sliding window approach to predict potential allergenicity of proteins.

It is done according to the current recommendations of the FAO/WHO Expert Consultation, as outlined in Codex alimentarius; however, this method generates false-positive and falsenegative hits, so it is advised by the FAO/WHO that the outcomes should be combined with other allergenicity assessment methods. APPEL (Allergen Protein Prediction E-Lab) (http://www.jing.cz3.nus.edu.sg/cgi-bin/APPEL) tool uses SVM to identify novel allergen proteins. This tool correctly classified 93 % of 229 allergens and 99.9 % of 6717 non-allergens. It is based on statistical method, and it has the potential to discover novel allergen proteins. EVALLER web server (http://www.slv.se/en-gb/Group1/Food-Safety/e-Testing-of-protein-allergenicity/) uses filtered length-adjusted allergen peptides (DFLAP) method (via ulfh@slv.se) to identify the potential allergen proteins. DFLAP extracts variable length allergen sequence fragments and employs SVM. EVALLER and APPEL servers assigned all calmodulins or calmodulin-like proteins as presumably non-allergens. But a conventional alignment approach (e.g., 35 % similarity over 80 amino acid segments) gives preference to find sequence similarity between input proteins and known allergens and puts abovementioned proteins in allergen category. These proteins are presumable non-allergenic homologues to the polcalcin family (members being potential allergens involved in pollen–pollen cross-sensitization). Tools, based on structural and physical characteristics, are useful

to identify potential cross-reacting proteins that may escape detection through sequence similarity method alone.

Applications of Immuno-informatics

The use of immunological databases and prediction software has become an important part of the scientific research as they allow us to predict the interaction of molecules involved in an immune response, thereby significantly shortening experimental procedure.

Reverse Engineering for Vaccine Design

Vaccines can be live attenuated whole pathogens, subunits, or epitope based. It is possible to design attenuated pathogens by removing virulence factors or reducing their metabolic capacity. These procedures can be done through computational design and discovery. Several in silico techniques have been developed to identify suitable vaccine candidates, principally proteins within pathogen genomes that have antigenic properties. Generally used vaccines are live attenuated or killed bacteria or viruses (examples include cholera, polio, measles). Thus, there is a concern about the safety of these vaccines; if they are incompletely attenuated or killed, they may revert their pathogenicity or cause undesirable immune reactions. On the other hand, synthetic peptides are considered as candidates for safe vaccines. Methods predicting immunogenic peptides could lead to rational vaccine design. Genome sequencing, comparative proteomics, and immuno-informatics tools are well developed to design new vaccines. “Reverse vaccinology,” a new concept, analyzes the entire genome to identify potentially antigenic extracellular proteins and thus helps in saving time and money. It was pioneered for *Neisseria meningitides* responsible for sepsis and meningococcal meningitides, and the vaccine type is conjugate based on capsular polysaccharide. These vaccines are available for pathogenic *N. meningitides A, C, Y, and W135*.

Microarray technique for vaccine design: Through microarray technology, it is easy to screen genes of various pathogens in different growth states and conditions for vaccine design. It reduces the number of genes useful for vaccine in a given genome. Signal peptides derived from genomic sequences, structural motifs, and immunogenicity are important for vaccine development.

Epitope-driven approaches for vaccine design: These are comparatively more useful as they have no lethal effect of the whole-protein vaccines. It may induce immune response against immunodominant epitopes. This kind of vaccine has a single start codon with an epitope which can be inserted consecutively in the construct. The prediction of promiscuous binding ligands is considered to be a prerequisite for the most subunit vaccine design strategies. It is originally named as “reverse immunogenetics” where T cell epitope mapping tools were employed to find new protein candidates for vaccines and diagnostic tests. Epitope-driven vaccine design allows the discovery of previously unknown and undescribed antigens and epitopes as vaccine candidates. The major disadvantage of the epitope-based approach is that algorithms may fail to predict all the relevant epitopes.

A web server, PEPVAC (Promiscuous EPitope-based VACcine) (http://immunax.dfci.harvard.edu/PEPVAC/), is optimized for the formulation of multi-epitope vaccines with broad population coverage. This optimization is accomplished through the prediction of peptides that bind to several HLA molecules with similar peptide-binding specificity.

Peptide-based vaccine design: Small peptides derived from epitopes are used as peptide-based vaccines. These peptides are recognized by MHC class I and thus boost the immune response. Three novel classes of methods have been described to predict MHC-binding peptides and a voting scheme to integrate them for improved results. The first method is based on quadratic programming applied to quantitative and qualitative data. Second method uses linear programming, and the third one considers sequence profiles obtained by clustering known epitopes to score candidate peptides. This method is found to be better than other sequence-based methods for finding the MHC binders.

Alignment-free approach for vaccine design: Some proteins have similar structure and biological properties, but they may lack sequence similarity. For these kinds of proteins, a new alignmentfree approach for antigen prediction has been proposed, which uses three datasets—each for bacteria, viruses, and tumours. The models were validated using leave-one-out cross-validation (LOO-CV) on the whole sets and by external validation using test sets and were implemented in a server called VaxiJen version 2.0 (http://www.ddg-pharmfac.net/vaxijen/). *DNA vaccines*: DNA vaccines produce cell-mediated and humoral immune response and are very useful in defending intracellular pathogens. It uses plasmid DNA, which contains a DNA sequence coding for an antigen and a promoter for gene expression in the mammalian cell. Plasmid DNA does not need a viral vector for delivery. Naked DNA is safe and can be used to sustain the expression of antigen in cells for longer periods of time than RNA or protein vaccines. The DNA delivers antigen as well as activates innate immunity and an adaptive immunity against cancer antigens. DyNAVacs (http://www.miracle.igib.res.in/dynavac/) incorporates different modules like codon optimization for heterologous expression of genes in bacteria, yeast, and plant, mapping restriction enzyme sites, primer design, Kozak sequence insertion, custom sequence insertion, and design of genes for gene therapy. The crucial question in deciding vaccine protocol is the vaccination schedule, i.e., is to decide whether the chronic protocol is able to give 100 % protection or shorter protocols could be applied. Thus, a mathematical model/simulator (SimTriplex) which describes the immune response activated by the triplex vaccine has been developed. Immunological prevention of cancer has been obtained in HER-2/neu transgenic mice using a vaccine that combines three different immune stimuli (triplex vaccine) that is repeatedly administered for the entire life-span of the host (chronic protocol). The software NERVE (http://www.bio.unipd.it/molbinfo) helps in designing subunit vaccines against bacterial pathogens. It combines automation with an exhaustive treatment of vaccine candidate selection task by implementing and integrating six different kinds of analyses. Xiang et al. developed a webbased database system, VIOLIN (Vaccine Investigation and Online Information Network) (http://www.violinet.org), which curates, stores, and analyzes published vaccine data. It contains four integrated literature mining and search programs, viz., Litsearch, Vaxpresso, Vaxmesh, and Vaxlert. They have

developed a web-based vaccine design system called Vaxign, which predicts possible vaccine targets. Major predicted features include subcellular location of a protein, transmembrane domain, adhesion probability, sequence conservation among genomes, sequence similarity to host (human or mouse) proteome, and epitope binding to MHC class I and class II. However, synthetic vaccine candidates must be tested experimentally to demonstrate their ability to generate neutralizing antibodies.

Immune System Modeling

The immune system can be seen as a parallel, information processing system that learns through examples, constantly adapts itself to new situations, and possesses a distributive memory for patterns. For theoretical immunology, immune system models and simulations can describe more insights into various interactions resulting in immunological phenomena. These models can test and find out the antigen–antibody interactions and immune responses for a particular antigen, in case of drug administration or testing of a vaccine candidate. Using visual modeling application described by Gong and Cai one can understand the adaptive immune system effectively. The hierarchical immune system consists of inherent immune tier, adaptive immune tier, and immune cell tier. It is designed and visualized with Java Applet technique for simulation. For further simulation purpose, the learning of the antibody is implemented through the evolutionary mechanism of the immune algorithm. ImmunoGrid (http://www.immunogrid.eu) and Virolab (http://www.virolab.org/) projects are working to simulate immune systems. ImmunoGrid tries to simulate immune processes by combining experiments and computational studies, while Virolab attempts to develop a virtual lab for infectious diseases by examining the genetic causes of human illnesses.
Exclusive computational approaches like mathematical modelling generate enormous amount of data, but there should be a balance between virtual and real experimental data. Computationally generated data needs to be formally tested and translated into real knowledge. Post-genomic era needs to exchange data from wet lab to simulation and vice versa. The model should be accurate, easy to use, and understandable to both model designers and biologists who can verify their hypothesis through in silico experiments.

Immuno-informatics for Cancer Diagnosis and Therapy

Antigen presentation plays a central role in the immune response and as a result also in immunotherapeutic methods like antitumour vaccination. There is a need to rapidly screen the antigens and to design specific types of expression constructs for immunotherapy of cancer. Competent immune responses to cancer are likely to be restricted to the immunome of a specific cancer, including the set of antigens that drive successful immune responses. However, it is still difficult to find the set of antigens that varies between different tumours. Antitumour vaccination takes advantage of in vivo processes, and it harnesses the full power of the immune system, unlike the more artificial ex vivo expansion of T cells. Changes in the cancer diagnosis and prevention are being supported by informatics. For example, the Cancer Biomedical Informatics Grid (caBIG) connects a network of 500 individuals and 50 institutions who share data and analyze tools to speed up the development of innovative approaches for the prevention and treatment of cancer. The 2005 database issue

of Nucleic Acids Research lists 14 cancer-related molecular databases, which mainly focus on cancer-related genes and gene expression. Listings of tumour antigens are also available. This list includes antigens that have defined T cell epitopes. Tumour-associated antigens (TAA) have played a vital role in both diagnosis and treatment of human carcinomas, such as prostate-specific antigen (PSA) in the diagnosis of prostate cancer. Despite this, the process of TAA identification has often been hampered by the complicated lab procedures. To fasten the process of tumour antigen discovery, and improve diagnosis and treatment of human carcinoma, a publicly available database Human Potential Tumour Associated Antigen (HPtaa) database (http://www.hptaa.org) has been established. Systems biology approaches target identification of a small number of antigens expressed by cancer cells that are suitable targets of immune responses against cancer. A proteomic mapping of in vivo targets for antibodies in lungs, and solid tumours in experimental animals define aminopeptidase-P and annexin A1 as targets of anticancer immune responses. Informatic methods have also been used for classification of tumours into subtypes, which supports decision making for the selection of therapeutic approaches; however, such applications in cancer immunology are yet to come.

Vaccine against tumours: Reliable predictions of immunogenic T cell epitope peptides are crucial for rational vaccine design and represent a key problem in immunoinformatics. Computational approaches have been developed to facilitate the process of epitope detection and show potential applications to the immunotherapeutic treatment of cancer. Epitope-driven vaccine design employs these bioinformatics algorithms to identify potential targets of vaccines against cancer. The development of epitope-based DNA vaccines and their antitumour effects in preclinical research against B cell lymphoma have been described. Most immunotherapeutic approaches work on the induction of antitumour CD8+ T cells, which exhibit cytolytic activity towards tumour cells expressing tumour-specific or tumour-associated Ags. But the immunization strategies that focus solely on CD8+ T cell immunity might prove to be insufficient because they will be unable to provide long-term protective immunity. It has been shown that the peptides predicted to bind MHC can elicit a tumour-killing cytotoxic T lymphocyte (CTL) response. Although CTLs have been found to be the key player in the generation of antitumour therapeutic effects, sometimes they also remain as suboptimal. CD4+ T cells are critical for the generation and maintenance of CTL response through providing cytokines or by major pathway, i.e., dendritic cell licensing. Class II MHC-bound epitopes activate CD4+ T cells and maintain effective CTL response that plays an important role in the antitumour response. CD4+ T cells determine the functional status of both innate and adaptive immune responses; thus, the inclusion of appropriate
CD4+ T cell epitopes may be essential for vaccine efficacy. Idiotypic immunoglobulin M (IgM) expressed by B cell lymphoma is a clonal marker and a tumour-specific antigen. Thus, it can be used as an immune target. Specific immunogenic epitopes identified from these tumour antigens can be used as vaccines to activate an immune response against tumour cells. Concerning to lymphoproliferative malignancies, tetanus toxin fragment C (TTFrC)-fusion vaccine design was able to activate anti-Id antibody responses and to suppress tumour growth in murine models as well as was effective in inducing CD8+ CTL in several tumour models.

Immuno-informatics and Systems Biology for Personalized Medicine

The idea to integrate immuno-informatics with systems biology approaches is for the better understanding of immune-related diseases at various systems levels. This integration can open the path of several translational studies for better clinical practices. The association between a disease and genetic variations is one of the most important aspects in pharmacogenomics and development of personalized medicine. Figure 2 shows the integration that leads to the development of personalized medicine. The information about allele frequencies of immune molecules in a human population is important as different patient subgroups can be identified with different vaccine or drug responses. This includes polymorphism information on HLA, cytokines, and killer-cell immuno-globulin like receptors (KIR). Thus, there is a scope for the development of optimized vaccines and drugs tailored to personalized prevention and treatment through the integration of systems biology and immunoinformatic.

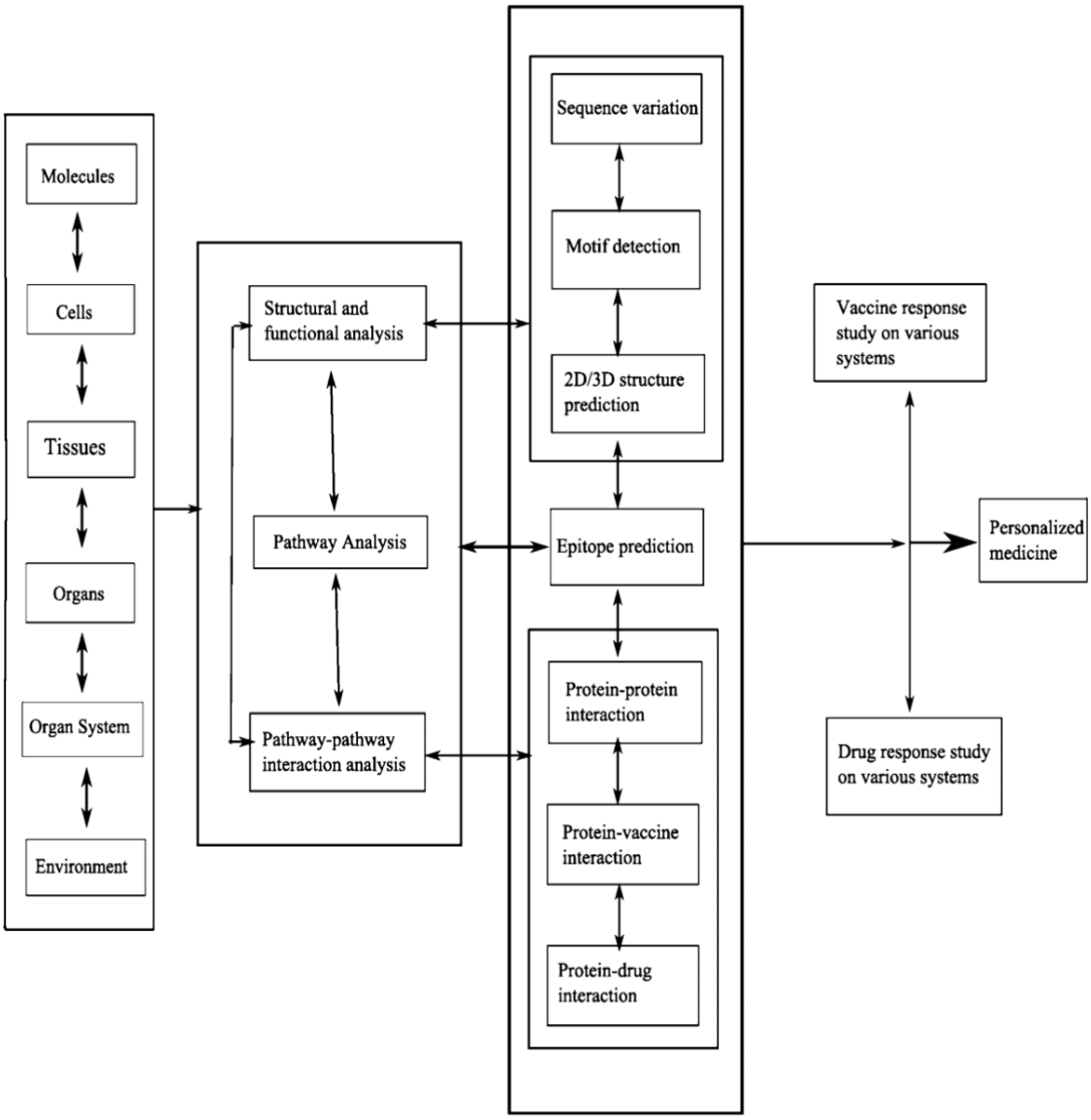

Fig. 2 An integration of immunoinformatics and systems biology, leading to the development of a personalized medicine

Conclusions and Discussions

High-throughput experimental techniques are combined with immuno-informatics, resulting in explosive growth of immunology. This is as similar as the event that has transformed genetics into genomics. Immuno-informatics may be placed at the junction point between experimental and computational approaches as it reduces time and cost involved in traditional study of immunology. This review considers useful online immunological databases, tools, and web servers and explores the application of immuno-informatics in various scientific domains with an emphasis on reverse vaccinology. Earlier approaches have some limitations in handling real data (nonlinear data). Machine-learning techniques can deal with nonlinear data. SVM (a statistical learning methodology) is a learning technique which supports continuous and categorical variables. SVM is better than ANN, as it attains global minimum and is capable of working with a smaller number of training patterns. Thus, both sequence characteristics and computational techniques should be integrated to acquire higher prediction accuracy. “Reverse vaccinology” is a revolution in immunology as it uses the whole spectrum of antigens. This helps in using pools of vaccine candidates which otherwise would be missed (because of poor or no in vitro experimental information or facing problem in culturing the specific pathogen). Recently, the prediction of promiscuous peptides (capable of binding to a wide array of MHC molecules) is being given much emphasis. Screening of large-scale pathogens and mapping of T cell epitopes allow identification of prime target of epitope-based T cell vaccine design.

Immuno-informatics models simulate the real behavior of immune system processes and thus help to get the kinetics of cells during immune responses. It is engineered in such a way that it can be studied and interpreted easily and can be rebuilt if new experimental data are introduced. These mathematical models remove the uncertainty of the systems as they are found to be closed to wet lab experiments. It leads to design the path for refinement and model the new experiments. But they cannot be directly compared to real biological data as they rely on assumptions only. There is no data for extended time spans available to validate the model. This limits the accuracy of the results. Currently models are designed in such a way that they simulate the biological data only over a fixed time period. It should have the ability to show the system’s changes over an extended time period for immune response in case of antigen attack or drug administration. This will reduce the necessity of experimental research. Drug response to a host’s immune system can be better studied through computational models. Effect of drug administration can be added to model the immune system to find the drug efficacy. Immune system/drug response study provides an idea about the dose composition, drug dosage duration, age of the patient, and other parameters. These modelling capabilities may lead to designing a drug, which can treat a disease without any side effects. Thus, the idea of integrating systems biology with immuno-informatics can lead to better clinical trials.

Reference:

Most of the libraries and websites found here are from research on the internet and work gathered over the years. Some other texts were found in the books below.

1. Immunoinformatics: an integrated scenario by Namrata Tomar and Rajat K. De (2010)

2. Cellular and Molecular Immunology, Updated Edition: With STUDENT CONSULT Online Access, 5e (Cellular and Molecular Immunology, Abbas) 5th Edition by Abul K. Abbas MBBS, Andrew H. H. Lichtman MD PhD