DEVELOPMENT OF IMMUNOINFORMATICS TOOLS

FOR VACCINE DESIGN

By

VARUN JAISWAL

A THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENT FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

IN

BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAHGHAT

APRIL, 2014

CERTIFICATE

This is to certify that the thesis entitled, **"Development of immunoinformatics tools for vaccine design"** which is being submitted by Varun Jaiswal for the award of degree of Doctor of Philosophy in Bioinformatics by the Jaypee University of Information Technology at Waknaghat, is a record of the candidate's own work, carried out by him under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Dr. Chittaranjan Rout
Assistant Professor (Senior Grade)
Dept. of Biotechnology and Bioinformatics
Jaypee University of Information Technology,
Waknaghat, Solan, H.P. India- 173234
Date: 22.10.2013

DECLARATION

I certify that

a. The work contained in this thesis is original and has been done by me under the guidance of my supervisor.
b. The work has not been submitted to any other organization for any degree or diploma.
c. Whenever, I have used materials (data, analysis, figures or text), I have given due credit by citing them in the text of the thesis.

Varun Jaiswal

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dr. Chittaranjan Rout for the continuous support of my Ph.D research. I am deeply indebted to my research supervisor for his constant encouragement, perennial interest, mentoring research support and untiring help during my difficult moments.

I emphatically express my loyal and venerable thanks to honorable Prof. S. K. Kak (Vice Chancellor) and Brig. (Retd.) Balbir Singh (Director, JUIT) for providing opportunity to pursue Doctorate Degree and advanced lab infrastructure to accomplish this scientific venture of my life. I am also very much thankful to Prof. T. S. Lamba for his support and cooperation. It gives me immense pleasure to express my gratitude to Prof. R.S. Chauhan (Dean and HOD, BT & BI) for his ever-smiling disposition coming to my rescue in solving my problems and suggestions that helped me in maintaining my confidence. I am grateful to DRDE and JUIT for providing scholarship during Ph.D research.

I express my sincere gratitude to teachers of BT & BI department, Prof. C. Tandon, Dr. Simran Tandon, Dr. Hemant Sood, Dr. Tirtha Raj Singh and Dr. Rahul Srivastava for their efforts to bring improvements in me both academically and personally.

Words are short to express my deep sense of gratitude towards my friends, colleagues, juniors and seniors, Namita, Tarun, Ravindra, Ashutosh, Anmol, Jeet, Aseem, Priya, Pallavi, Asha, Swapnil, Charu, Vivek, Madan, Arun, Dr. Jatin, Dr. Surjeet and Dr. Sree Krishna. Their presence and imperative help made my research trouble-free and enjoyable.

I am grateful to my parents, elder brother and sister-in-law, nephew (Yuvraj) and niece (Tanishka). Their love provided me inspiration and was my driving force. I owe them everything and wish I could show them just how much I love and appreciate them.

Finally, I bow my head before almighty for his endless blessings and giving me patience, encouragement and strength to achieve this academic milestone.

Varun Jaiswal

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AAP	Amino Acid Pair
ALYS	Annual Life Year Saved
ANN	Artificial Neural Network
BCE	B-Cell Epitope
BEID	B-Cell Epitope Interaction Database
BW	Bio-Weapon
BWB	Bio-Weapon Bacteria
CBP	Choline Binding Protein
CDC	Center for Disease Control and Prevention
CGI	Common Gateway Interface
CSC	Cumulative Strain Coverage
CSV	Comma Separated Values
CTL	Cytotoxic T Lymphocye
DALYs	Disability-Adjusted Life Years Saved
DDBJ	DNA Data Bank of Japan
EBI	European Bioinformatics Institute
ECE	Epitope Combination Explorer
EIR	Epitope Information Resource
FAO	Food and Agriculture Organization
FBP	Fibronectin-Binding Protein
FPIA	Fusion Proteins for Immune Applications
FPIA	Fusion Proteins for Immune Applications
FSA	Forward Selection Algorithm
GI	Genome Accession Id
GISN	Global Influenza Surveillance Network
HA	Hemagglutinin
Hib	*Haemophilus influenzae* type B
HIV	Human Immune-Deficiency Virus
HLA	Human Leukocyte Antigen
HMMs	Hidden Markov Models

HPA	Human Platelet Antigens
HBsAg	Hepatitis B Virus Surface Antigen
HTML	Hypertext Markup Language
IEDB	Immune Epitope Database
IG	Immunoglobulin
IgSF	Immunoglobulin SuperFamily
IRD	Influenza Research Database
ISC	Individual Strain Coverage
LAV	Live-Attenuated Vaccine
LYS	Life Years Saved
M1	Matrix Protein 1
M2	Matrix Protein 2
MHC	Major Histocompatibility Complex
MhSF	MH superfamily
MMdb	MUGEN Mouse Database
NA	Neuraminidase
NCBI	National Center for Biotechnology Information
NERVE	New Enhanced Reverse Vaccinology Environment
NICs	National Influenza Centres
NP	Nucleoprotein
NS1	Non Structural Protein 1
NS2	Non Structural Protein 2
OmpA	Outer-Membrane Protein A
PA	Protective Antigen
PBP	Penicillin-Binding Protein
PLS-DA	Partial Least Square Discriminant Analysis
PSSM	Position Specific Scoring Matrices
PVC	Protein Vaccine Candidates
PVCs	Protein Vaccine Candidates
RDT	Recombinant DNA Technology
RF	Random Forest
RV	Reverse Vaccinology
SBPs	Solute Binding Proteins

TANTIGEN	Tumor T Cell Antigen Database
TBP	Transferrin-Binding Protein
TCE	T-Cell Epitope
Th	T-Helper
TIV	Trivalent Inactivated Vaccine
TR	T-Cell Receptor
UIV	Universal Influenza Vaccine
UNICEF	United Nations Children's Fund
URL	Uniform Resource Locator
VLPs	Virus Like Particles
WHO	World Health Organization

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Vaccine: origin and evolution

Infectious diseases are one of the leading causes of mortality. Prevention is always better than cure and the vaccines are still the most successful preventive measures against infectious diseases (Ehreth, 2003). Although the first use of vaccine was much ancient and many of them were started independently in Africa, China and India (Gross and Sepkowitz, 1998). Vaccine was formally introduced in medical practice by Edward Jenner through his landmark experiment in which he had shown that infected material from cow containing 'cowpox' can be used to protect human from smallpox (Riedel, 2005). A century after Edward Jenner's smallpox vaccine, Louis Pasteur proposed basic rules of vaccinology based on germ theory of disease and developed the first vaccines against anthrax and rabies (Geison, 1978). Pasteur proposed that vaccine can be developed through isolation, inactivation and injection of pathogenic microorganism.

After the advances in tissue culture techniques of viruses which were introduced by Hugh and Mary Maitland in 1937, Max Theiler used different types of tissue cultures for the cultivation of yellow fever virus. These tissue culture techniques helped in attenuation of viral growth in human and produced 17D attenuated virus strain for live attenuated vaccines (Theiler and Smith, 1937). Later in 1951, Max Theiler got Noble prize for the development of yellow fever virus vaccine. Similarly, cultivation of viruses in embryonated hens' eggs by Ernest William Goodpasture and his colleagues at Vanderbilt University was the basis for development of first influenza vaccine. Ernest Goodpasture found that several (more than 30) viruses could be propagated in the chorioallantoic membrane which surrounded the chick embryo. Subsequent researchers used hen's eggs to cultivate influenza viruses which were later used successfully to develop influenza vaccine (Plotkin and Plotkin, 2004).

In a breakthrough in cell culture technique, John Enders and his colleagues at the Children's Hospital, Boston successfully cultivated the poliovirus in human embyonic cell culture in 1948 (Enders *et al*., 1949). This cultivation of poliovirus greatly facilitated the development of first polio vaccine in 1952 by Jonas Salk. Later, John Enders and his colleagues were awarded with the Noble prize in 1954 for their work on cultivation of poliovirus in human embryonic cell culture. In 1963, the Ender also developed measles vaccine which was the live attenuated strain (edmonston strain) of measles virus (Hilleman, 1992). In last two centuries, after the formal discovery of vaccine by Edward Jenner, the revolution in invention of vaccines against different pathogenic organism was driven mainly by new findings and advancement in cultivation techniques. The germ theory of disease was

the basis of principle proposed by Louis Pasteur for preparing vaccines. Further, advancement in culture techniques such as tissue culture, cell culture, virus culture in eggs, etc. had revolutionized the vaccine design, research and production in 20th century.

At the end of twentieth century, the most successful ways of making vaccines which were based on Pasteur's principle: isolation, inactivation and injection of disease causing pathogen. New techniques were the need of the hour to discover/invent vaccines against the diseases which were not feasible through above simple principle. New approaches such as recombinant DNA technology (RDT), protein engineering, chemical conjugation of protein and polysaccharides, and use of novel adjutants have opened the new era of vaccines research. Before discovery of the RDT, the manufacturing of hepatitis B virus vaccine was limited, as antigen required for the production of vaccine could only be recovered from patients infected with hepatitis B virus. But RDT assisted to produce recombinant proteins of hepatitis B virus on a large scale (Rappuoli, 2007). Chemical conjugation of proteins and polysaccharides in vaccine provided both cellular and humoral immune responses to prevent infection. The conjugation methods have opened the way to development of several effective vaccines against pathogens such as *Haemophilus influenzae* (Morris *et al*., 2008)*, Neisseria meningitidis* (serogroup A, C, Y and W135), *Streptococcus pneumonia* (Sinha *et al*., 2007), etc. Furthermore the prevailing resource for vaccine design, the availability of genomes (host and pathogens), came out when genome of *H. influenza* was published in late twentieth century (Fleischmann *et al*., 1995). With the accessibility of genome, virtually all possible proteins of an organism were available for evaluation of their potential as vaccine candidates. Vaccine development has been coupled with advancement of new techniques and associated discoveries. The milestone breakthroughs associated with events in vaccine research and developments are summarized in Table 1.1. Overall new technologies such as computational based methods (like reverse vaccinology), synthesis of customized DNA sequences, expression of antigen in a selected host, reconstruction of viral genomes, generation of recombinant vectors (viral or bacterial) and codon optimization for expression of antigens formulate the section of synthetic biology which used in vaccine design and development (Kindsmülle and Wagner, 2011). *De novo* synthesis of DNA and RNA through synthetic biology eliminates time and efforts required for pathogen cultivation and will be leap forward towards production. Synthetic Genomics Vaccines Inc. (SGVI) focusing in synthetic biology for fast production of influenza vaccines which are required to be updated in every season and up-gradation of vaccine must done in minimum time (Kindsmülle and Wagner, 2011).

Table 1.1: Major breakthroughs that influenced vaccine development

Events	Year	Outcome	Reference
Formal discovery of vaccine	1796	Small-pox vaccine	(Riedel, 2005)
Pasteur principle of vaccinology (Isolation, inactivation and injection of pathogenic microorganism)	1880	First vaccines against the diseases anthrax and rabies	(Pasteur, 1880)
Culture of viruses in hens egg	1931	Influenza vaccine	(Plotkin and Plotkin, 2004)
Advancement in tissue culture of viruses	1937	Yellow fever virus vaccine	(Theiler and Smith, 1937)
Cell culture of viruses in human embryonic cells	1948	Polio vaccine	(Enders *et al*., 1949)
Recombinant DNA technology	1972	Production of recombinant Hepatitis B vaccine (Subunit vaccine)	(Jackson *et al*., 1972)
Reverse vaccinology	2000	*Neisseria meningitidis* serogroup B vaccine	(Rappuoli, 2000)

1.2 Importance of vaccines

Since more than 200 years, vaccines are the most successful and cost effective preventive measures against infectious diseases (Ehreth, 2003). These are instrumental in preventing humans and animals from microbial infections. Vaccines were also the main weapons in eradication of smallpox and rinderpest diseases (Normile, 2008). Vaccines also set the disease poliomyelitis near to eradication and an array of other diseases partly under control due to vaccination. According to the UNICEF, vaccines worldwide save approximately nine million lives annually (http://www.unicef.org/pon96/hevaccin.htm). In terms of annual life year saved (ALYS) and disability-adjusted life years saved (DALYs), the vaccines contributed huge benefit through eradication of smallpox, and control of polio, measles and tetanus as shown in Table 1.2 (Ehreth, 2003).

Potential threat of bioweapons also increases the role for vaccines in protection against deadly infectious disease caused by bioweapons. Now vaccines are available for some bacterial bioweapons such as *Bacillus anthracis*, *Francisella tularensis* and *Yesinia pestis*. But these vaccines are not highly effective and are having side-effects. However, these vaccines are administered to frontlines workers such as military, doctors and other people in an eminent war like situation. The major global viral diseases prevented through vaccines are adenovirus-based diseases, hepatitis A, hepatitis B, human papillomavirus, influenza,

Japanese encephalitis, measles, mumps, polio, rabies, rotavirus diarrhea, rubella, smallpox, tick-borne encephalitis, Varicella zoster, yellow fever. Similarly, important global bacterial diseases prevented or partly mitigated through vaccines are cholera, diphtheria, meningococcal meningitis, pneumococcal pneumonia, typhoid fever, tuberculosis, tetanus, *Haemophilus influenzae* and plague (Koff *et al*., 2013). Currently, more than two thousand (2162) licensed vaccines are available against 124 diseases. These vaccines are against 9 Gram-positive bacteria, 35 Gram-negative bacteria, 72 viruses, 8 parasites and cancer (http://www.violinet.org) (Xiang *et al*., 2008).

Table 1.2: Benefits disease eradication or control by vaccination in terms of annual life years saved (LYS) and disability-adjusted life years saved (DALYs) (Ehreth, 2003)

S. No.	Disease	LYS	DALY
1.	Smallpox	5,000,000	NA
2.	Polio	35,750,000	1,725,000
3.	Measles	71,500,000	29,838,000
4.	Tetanus	56,030,000	12,020,000

1.3 Types of vaccines

First successful vaccine was the live smallpox vaccine and currently various types of vaccines are used against different infectious diseases. Vaccines in use are categorized as live attenuated-, inactivated-, subunit-, conjugate- and toxoid-vaccines but a few types of vaccines are still in research or evaluation stages, like DNA vaccines, recombinant vector based vaccines, peptide-based (epitope) vaccines, etc. These distinctive types of vaccines also differ in terms of their diverse potential for protections and safety issues.

1.3.1 *Live attenuated vaccines (LAVs)*

A live-attenuated vaccine (LAV) is produced by reducing the virulence of a microbe, but still keeping it live. In attenuation process, the microbe (the infectious agent) is altered in a manner so that it becomes less or non-virulent, thereby not causing any diseases in host but still it preserves the antigenic elements to produce immune response. LAV is the closest thing to a natural infection as compare to other types of vaccines. This kind of vaccine elicits strong humoral as well as cellular immune responses and generally provides lifelong immunity with only few vaccinations. Although LAV has advantages over other types of

vaccines as it is more natural but safety issues are associated with it. These LAVs are not administered to immune-compromised individuals. Sometime emergence of novel pathogenic strains through mutation or any other means that has risk to regain virulence of microbe makes the LAV ineffective. Other limitations with LAVs are storage and transportation that generally require refrigerated condition to retain potency of vaccine. Overseas shipping and storage especially to developing countries where health system lacks refrigeration facility is still the issues associated with use of this kind of vaccine. The LAVs against several viruses are developed through cultivation of viruses in different types of cultures, and continuous development of new cultures resulted in decrease of the virulence of the viruses. Small genome size makes virus the simpler microbe so researcher can easily control the characteristic of viruses for LAVs as compared to bacteria. Currently, LAVs are available against diseases such as measles, mumps, rubella, influenza, chicken pox, polio, diarrhea caused by rotavirus, yellow fever, rabies, varicella, etc.

1.3.2 ***Inactivated vaccines***

Inactivated vaccine is produced by killing of the disease causing organism through chemical, temperature, or radiations. These vaccines are safer than live attenuated vaccines (LAVs). These vaccines are killed microbes therefore reversion of virulence is not possible. Killed vaccines generally do not require refrigeration so these types of vaccines can be stored and transported in freezed/dried form. Drawback with these inactivated vaccines is that they produce weaker immune response as compared to LAVs and require booster doses time to time.

1.3.3 ***Subunit vaccines***

Subunit vaccines contain only the part of microbe (antigen) which can provide immune response in host. Sometime subunit vaccine contains only epitopes (only part of antigens which recognized by antibodies or T-cell receptors). Adverse reactions against subunit vaccine are low because it contains only part of pathogen important for immunity instead of all constituents of microbe. Finding antigen from microbe which can be used as subunit vaccine requires several expertise (microbiology and immunology) and resources (cost and time). If antigen is known for subunit vaccine then manufacturing of antigen may be done through recombinant DNA technology (RDT), and the developed vaccine is known as recombinant subunit vaccine. This method has been used to develop vaccine against Hepatitis

B virus (Rappuoli, 2007). Conventional methods used for finding antigens had its limitation, and these methods also required huge resources and time. Computational methods have been assisting subunit vaccine development by increasing its effectiveness as well as reducing required resources and time. Still more accurate computational methods are required to boost subunit vaccine development (Rappuoli, 2000).

1.3.4 ***Conjugate vaccines***

Conjugate vaccines have both polysaccharide and protein as their constituents which are covalently attached. Conjugate vaccine is special type of subunit vaccine which has potential to provide both T-cell and antibody mediated immune responses. Several pathogenic bacteria have outer coating formed by polysaccharides which are immunogenic. These polysaccharides in conjugation with protein are used to develop such vaccines which are expected to provide better immune responses. Vaccines that protect against *Haemophilus influenzae* type B (Hib) and meningococcal disease (MCV4) are conjugate vaccines (Einhorn *et al*., 1986).

1.3.5 ***Toxoid vaccines***

Toxoid vaccines are the inactivated toxins of microbes. In case of some microbes, secreted toxins or harmful chemical are the major factors for illness, therefore, toxoid vaccines are used to get protection from these toxins which are secreted by microbes. These vaccines are generally prepared through inactivation of toxins using formalin or any other means. Generally, toxoid vaccines stimulate antibodies as their immunogenic units are similar to natural toxins. Since the structural part of toxin important for immunogenic response is preserved during toxoid preparation, administration of toxoid vaccines assists immune system in learning how to handle the natural toxin. Vaccine used against diphtheria and tetanus are toxoid vaccines (http://www.fda.gov/BiologicsBloodVaccines/Vaccines/ApprovedProducts/ucm094012.htm).

1.3.6 ***DNA vaccines***

DNA vaccines are genetically engineered DNA to produce protective immunological response. A small and circular DNA (like plasmid) is genetically engineered to produce specific antigenic proteins from a particular pathogen. When DNA vaccines are introduced to host, some of the host cells take up that DNA vaccine and use host cell machinery to

synthesize the pathogen's proteins which are subsequently secreted and/or displayed on the surface of host cell for immune recognition and responses. DNA vaccines can provide strong antibody response to secreted pathogen's proteins (antigens) as well as strong cellular immune response to antigen displayed on cell surface. There is no risk of disease with DNA vaccines because these types of vaccines contain only few proteins of pathogen in their DNA.

1.3.7 ***Recombinant vector vaccines***

These are DNA vaccines which use bacterial or viral vector to introduce microbe's DNA (codes for antigen) to host. In nature, viruses are known to inject their DNA into the host cell so the virus DNA containing gene(s) (other than the virus genes) of microbe's protein(s) which can be used as vector. Bacteria can also be used as vector and in that case the attenuated or harmless bacteria are inserted with DNA from pathogenic microbes. Bacterial vector displays pathogenic microbe's proteins from inserted DNA on their surface which induce immune responses. Both viral and bacterial recombinant vector vaccines such as recombinant attenuated Salmonella vaccine vectors (Curtiss III *et al*., 2010), recombinant poxviruses as mucosal vaccine vectors (Gherardi and Esteban, 2005) and recombinant vectors as influenza vaccines (Kopecky-Bromberg and Palese, 2009) are in research stages of vaccine development. Similarly, this method is under process for the development of vaccines against several challenging diseases including AIDS, rabies and influenza (Chapman *et al*., 2010; Chen *et al*., 2013; Sedova *et al*., 2012). These recombinant vector vaccines are similar to DNA vaccines except for use of recombinant virus or bacteria as vector. Although these vaccines are still in different stages of evaluation (clinical trials), but current researches have highlighted their potential in inducing both humoral and cellular immune responses (Barouch *et al*., 2013).

1.3.8 ***Virus-like particles (VLPs) vaccines***

Virus-like particles (VLPs) are multi-proteins which contain organization and conformation like native viruses without viral genome. VLPs contain repetitive viral surface proteins which exhibit conformational viral epitopes that can produce both T-cell and B-cell immune responses (Roldão *et al*., 2010; Jeong *et al*., 2004). VLPs are helpful means for the development of safer vaccines alternative to attenuated viruses because they lack genetic material. Hepatitis B virus surface antigen (HBsAg) VLP was the first licensed VLP vaccine (Michel *et al*., 2010). Currently, several other virus pathogens including hepatitis C virus,

chikungunya virus, influenza virus, ebola virus (EBOV), marburg virus (MARV) and HIV being targeted through VLPs vaccines are in research stages (Akahata *et al*., 2010; Jeong *et al*., 2004; Doan *et al*., 2005; Quan *et al*., 2007; Warfield *et al*., 2011).

1.4 Gaps and bridges in vaccine design

Although vaccines are available against diseases caused by sizeable number of pathogenic organisms but preventive measures (vaccines) are not available or ineffective against infection caused by large numbers of pathogenic microorganisms. Major global diseases for which vaccines are neither effective nor available include bacterial diseases (tuberculosis, urinary tract infections, campylobacter, chlamydia, gastrointestinal ulcers (*Helicobacter pylori*), shigella, streptococcus group A and streptococcus group B), viral diseases (dengue, influenza (universal influenza vaccine not available), cytomegalovirus, epstein-Barr (mononucleosis), hepatitis C, herpes Simplex, HIV, respiratory syncytial virus and rhinovirus) and parasite diseases (leishmaniasis, malaria, and schistosomiasis) (Koff *et al*., 2013). These diseases exert great challenge for healthcare researcher to design effective vaccines against them. Potential use of infectious diseases in bioweapon (BW) attack also raises the concern for effective vaccines against the bioweapons. Intelligence estimated that BW threat is greater than the nuclear attack because of the ease in dissemination of deadly pathogens (D'Agostino and Martin, 2009). On other hand, there are no vaccines against BWs such as Brucella, Salmonella and Burkholderia infections. Although vaccines are available for some bioweapon bacteria (BWB) but they have their limitations. Anthrax vaccines have serious side effects and those vaccines required yearly boosters (Weiss *et al*., 2007). Vaccine against *Francisella tularensis* is not fully licensed and data about efficacy of plague vaccine is not available (Jefferson *et al*., 1998).

Emergence of new and old virulent pathogenic organisms and acquisition of drug resistance by pathogens along with the threat on potential use of bioweapon (BW) calls for new and effective vaccines to prevent life-threatening diseases. But there were two major gaps in knowledge related to vaccine design: antigen variability and T cell immunity (types of immune response) (Figure 1.1). The role of these gaps in vaccine design can be delineated from graph between antigenic variability and types of immune response (Rappuoli, 2007). These gaps hindered the development of effective vaccines against the challenging diseases (Rappuoli, 2007). Different colours in graph indicated the vaccine status of pathogen: red colour portion of graph designates the pathogens for which licensed vaccines are available;

the orange area in the graph represents the pathogens for which vaccine is not available but discovery of vaccine against these pathogens is not challenging and RV technique can used to discover vaccines; and yellow colour segment covers the pathogen against which vaccine design is challenging, and structural vaccinology and immune-modeling are essential to develop vaccines against such diseases. Till date, the licensed vaccines either have no or less antigenic variability and they are also controlled mainly through antibody-mediated immune responses of host (Figure 1.1 (red region)). While pathogen against which vaccine design is challenging and no vaccines (Figure 1.1 (yellow region)) are available having high antigenic variability and they also require T-cell immune response for protection. Considering the graph, though antigenic variability can be addressed through the search in whole repertoire of protective antigenic proteins from a genomes but knowledge gap in T-cell immunity is difficult to measure. Discovery of antigen providing appropriate T-cell immune response could bridge the second knowledge gap in vaccine design (Rappuoli, 2007).

Finding protective antigens through whole genomes is termed as reverse vaccinology (RV) which is considered as first bridge for the knowledge gap in vaccine design. RV has provided great flexibility to explore whole proteome of pathogens without even culturing them. This technique had justified its importance in very first occasion when attempts to develop vaccines against *Neisseria meningitidis* serogroup B were failed due cross reactivity of capsular polysaccharide based vaccine with human tissues (Häyrinen *et al*., 1995) and the variability of outer membrane proteins (Poolman, 1995). The immense potential of RV method was widely recognized and through this techniques dozens of human and animal pathogens including *Neisseria meningitidis* (Pizza *et al*., 2000), *Helicobacter pylori* (Chakravarti *et al*., 2000), *Streptococcus pneumoniae* (Wizemann *et al*., 2001), *Porphyromonas gingivalis* (Ross *et al*., 2001), *Chlamydia pneumoniae* (Montigiani *et al*., 2002), *Bacillus anthracis* (Ariel *et al*., 2002), *Leptospira interrogans* (Yang *et al*., 2006), *Streptococcus suis* (Liu *et al*., 2009), extraintestinal pathogenic *Escherichia coli* (Moriel *et al*., 2010), *Echinococcus granulosus* (Gan *et al*., 2010), *Brachyspira hyodysenteriae* (Gan *et al*., 2010), Cryptosporidium species (Manque *et al*., 2011), *Haemophilus parasuis* (Hong *et al*., 2011), *Leptospira borgpetersenii* (Murray *et al*., 2012), *Pasteurella multocida* (Hatfaludi *et al*., 2012), *Edwardsiella tarda* (Zhang *et al*., 2012a), Leishmania spp (John *et al*., 2012), Leptospira serovars (Umamaheswari *et al*., 2012), *Ehrlichia ruminantium* (Liebenberg *et al*., 2012), *Vibrio cholera* (Barh *et al*., 2013), *Staphylococcus aureus* (Oprea and Antohe, 2013) and *Brucella melitensis* (Gomez *et al*., 2013) were targeted.

The second bridge in vaccine design is the antigenic variability which is more prominent in case of viruses. The variability of antigen can be addressed through using conserved vaccine candidates which can be determined from genome sequences of different strains of a pathogen (Figure 1.1) (Rappuoli, 2007). As emanated from Figure 1.1, immunoinformatics and comparative genomics are essential to develop vaccines against highly variable pathogenic microbes. In case of viruses, the variation in sequences and structures of surface proteins is the main stumbling block in vaccine design. High variability is one of the most important reasons behind non-availability or less effective universal vaccines for several viruses such as influenza A virus, HIV, rotavirus, etc. (Fiore *et al*., 2009; Johnston and Fauci, 2008; Kirkwood, 2010). The challenge before scientific community is to develop universal influenza vaccine (UIV) which overcomes the problems of global influenza strains information collection in each season for vaccine formulation. Sometime the virus uses genetic shift and drift to produce novel pandemic or endemic strains which results in significant number of deaths despite seasonal updated vaccines. Use of functionally important conserved antigenic peptides (B-cell and T-cell epitopes) is expected to ameliorate the impact of pandemic and endemic strains. Computational resources assist to bridge the gaps involved in vaccine design.

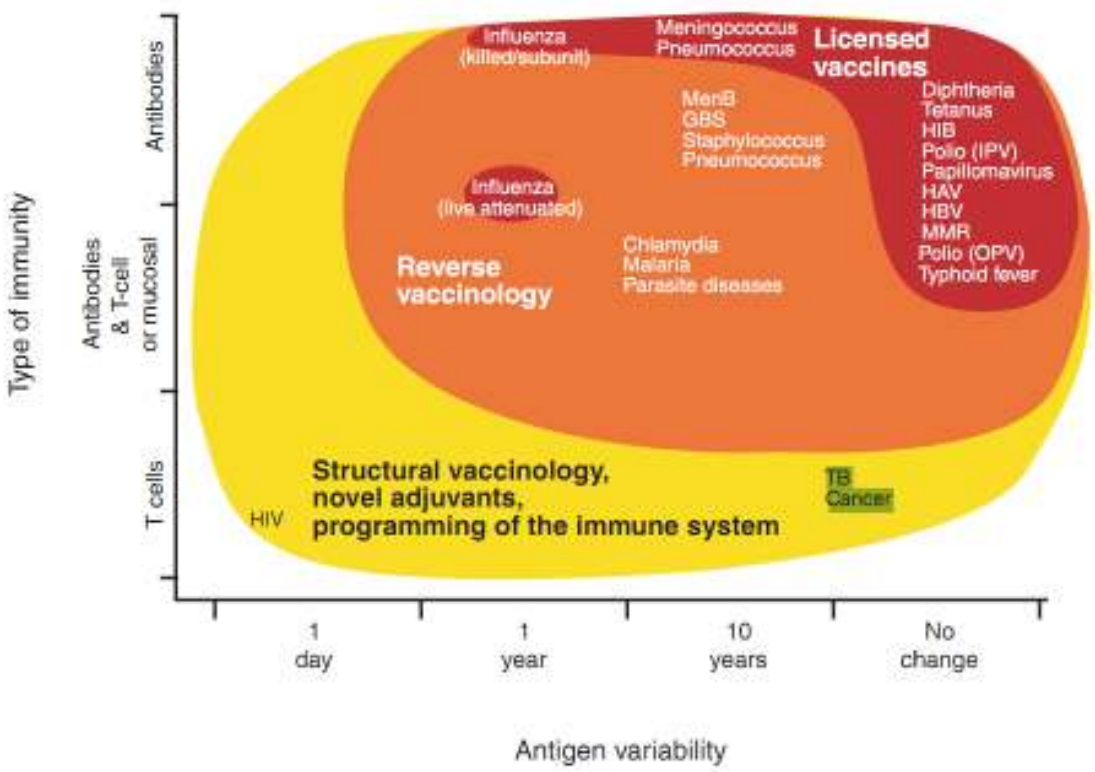


Figure 1.1: Graphical representation of antigen variability and immune response against pathogens. Current vaccines are mostly in the upper right red quadrant. Reverse vaccinology, which finds antigens able to induce protective antibodies, may extend considerably the area covered by vaccines (orange segment). The most difficult challenges are in the lower and left part of the graph, where antigens are extremely variable and protection relies only on T cells (Rappuoli, 2007).

Vaccines are developed mainly against those pathogens which are entirely prevented through suitable amounts of immunoglobulin antibodies in the serum. Protection of vaccine through T-cell response has strong theoretical background but there is no method to quantify protection of vaccine other than antibodies. For example, Sabin oral polio vaccine induces concentrations of serum antibodies that are lower than those induced by the killed Salk vaccine (Rappuoli, 2007). Effectiveness of Sabin vaccine is also attributed to protection beyond antibodies but we do not know how to measure that extra protection. There are other examples where T cells are likely to contribute to immune protection. Therefore, understanding of T-cell immunity may lead to the eradication of or protection from the most difficult diseases such as HIV, cancer and TB as presented in quadrant with yellow in Figure 1.1 (Rappuoli, 2007).

1.5 Web resources for vaccine design

High throughput experiments in molecular biology and immunology have been producing escalating quantity of data. Computational resources are necessary to store and analyze these huge amounts of biological data. To consider breadth and importance of topics in immunology, a new branch 'immunoinformatics' has evolved to assist vaccine design and development (Cohen *et al*., 2010). This branch uses genomics, proteomics, immunological methods, mathematics, information technology and computer science to bridge immunology and informatics (Petrovsky *et al*., 2003). Several databases primarily related to immunology and pathogen, web servers, and tools were developed to facilitate vaccine design. Immune epitope database (IEDB), The International ImMunoGeneTics Information System (IMGT) and Immuno Polymorphism Database (IPD) are the few most important and highly integrated immunoinformatics resources used for immunology and vaccine design research.

1.5.1 ***Immune epitope Database (IEDB)***

IEDB is the web portal containing data related to antibodies and epitopes for humans, non-human primates, rodents, and other animal species. It also has workbench and tools for analysis. The main classes of information within IEDB and their association which used in implementation of web resources are represented in Figure 1.2. Currently IEDB stores 104443 peptidic epitopes, 1931 non-peptidic epitopes, 208773 T-Cell Assays data, 161148 B-Cell Assays data, 8106 MHC ligand elution assays, 247807 MHC binding assays, 653 restricting MHC alleles, epitopes are form 3054 source organisms and 15196 references are associated with the existing data.

IEDB also has collections of several computational tools which assist vaccine design. These tools are available in three broad categories: T-cell tools, B-cell tools and analysis tools. T-cell prediction tools include peptide binding to MHC class I molecules: These tools take an amino acid sequence or set of sequences as input and determine each subsequence's ability for binding to specific MHC class I molecule. Similarly, T-cell prediction tools bind to MHC class II molecules: These types of tools employ different computational methods to predict MHC Class II epitopes, including a consensus approach which combines NN-align, SMM-align and combinatorial library methods. T cell epitopes - processing prediction tools: These tools predict epitope candidates based upon the processing of peptides in cell and the used processing techniques are proteosomal cleavage, TAP transport, MHC binding. T cell Epitopes - immunogenicity prediction tools: This tool predicts the relative ability of a

peptide/MHC complex to elicit an immune response. B-cell tools include linear B-cell epitope prediction tools, discontinuous B-cell epitope prediction tools and epitope prediction based upon structural protrusion tools. Analysis tools include population coverage, epitope conservancy analysis, epitope cluster analysis and homology mapping tools. All these tools are available at: (http://tools.immuneepitope.org/main/) (Vita *et al*., 2010).

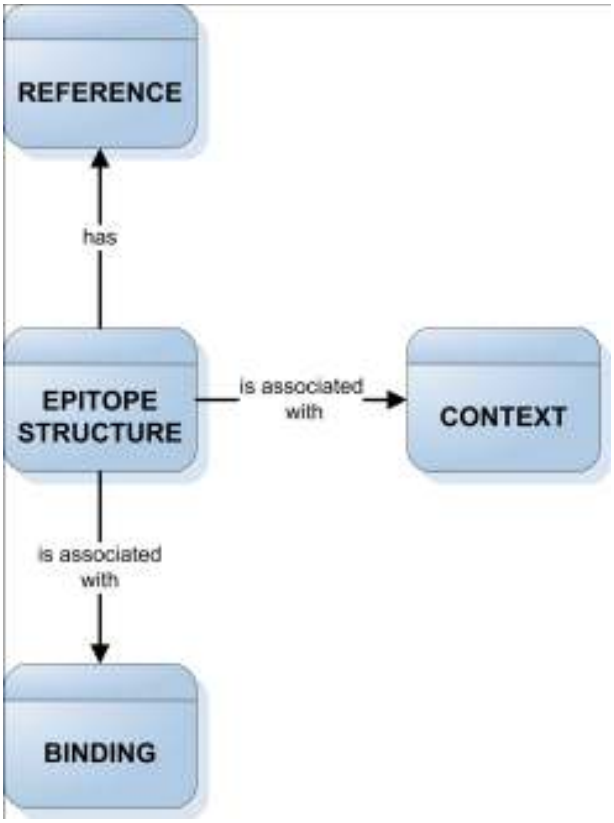


Figure 1.2: The main classes and their association implemented in IEDB web resource (Vita *et al*., 2010).

1.5.2 ***International ImMunoGeneTics Information System (IMGT)***

IMGT resource is the global reference in immunogenetics and immunoinformatics with a high-quality integrated knowledge resource specialized in the immunoglobulins (IGs) antibodies, T cell receptors (TR), major histocompatibility complex (MHC) of human and other vertebrate species, immunoglobulin superfamily (IgSF), MH superfamily (MhSF) and related proteins of the immune system (RPI) of vertebrates and invertebrates. This resource also provides a common access to sequence, genome and structure of Immunogenetics data. This organization works in collaboration with EBI DDBJ and NCBI, and has sequence databases, genome database, structure database, monoclonal antibodies database, web

resources and interactive tools. The complete overview of IMGT is provided in Figure 1.3. Currently the IMGT consists of seven databases: IMGT/LIGM-DB contains nucleotide sequences of IG and TR from 335 species (172,350 entries); IMGT/MH-DB contains sequences of the human MHC (HLA); IMGT/PRIMER-DB contains oligonucleotides (primers) of IG and TR from 11 species (1,864 entries); IMGT/CLL-DB contains IG sequences from chronic lymphocytic leukemia; IMGT/GENE-DB contains international nomenclature for IG and TR genes from human, mouse, rat and rabbit (3,107 genes, 4,722 alleles); IMGT/3Dstructure-DB and IMGT/2Dstructure-DB contains 3D structures of IG antibodies, TR, MH and RPI (2,802 entries); and IMGT/mAb-DB contains monoclonal antibodies (IG, mAb) and fusion proteins for immune applications (FPIA) (456 entries). IMGT resource also hosts tools for analysis in stored or user given data.

The IMGT contains tools for analysis: IMGT/V-QUEST software performs sequence alignment for IG and TR; IMGT/HighV-QUEST is used for high-throughput analysis on next generation sequencing (NGS) data of IG and TR; IMGT/JunctionAnalysis is used for human and mouse IG and TR junction analysis; IMGT/Allele-Align is used for alignment two sequences; IMGT/PhyloGene is used for phylogenetic analysis of IMGT standardized reference sequences; IMGT/DomainDisplay is used for the display of amino acid sequences from the IMGT domain directory; IMGT/GeneInfo provides information on data resulting from the mechanisms of V-J and V-D-J gene rearrangements in the T cell receptor (TR) loci of *Homo sapiens* and *Mus musculus*; IMGT/GeneFrequency provides a graphical representation of the numbers of cDNA and genomic IMGT/LIGMDB sequences containing the rearranged immunoglobulin (IG) and T cell receptor (TR) genes; and MGT/DomainGapAlign is used to create gaps in user provided amino acid sequence according to the IMGT unique numbering, for V-REGION or C-DOMAIN. IMGT/Collier-de-Perles is used to draw four types of domains: variable (V) domain, constant (C) domain, scavenger (S) domain of the immunoglobulin (IG), T cell receptor (TR) and other members of the immunoglobulin superfamily (IgSF), and groove (G) domain of the major histocompatibility complex (MHC) and other members of the MHC superfamily (MhcSF). IMGT/DomainSuperimpose is used for superimposing two IMGT domain 3D structures, and IMGT/StructuralQuery is used for three-dimensional structure analysis including chain details and contact analysis at different levels.

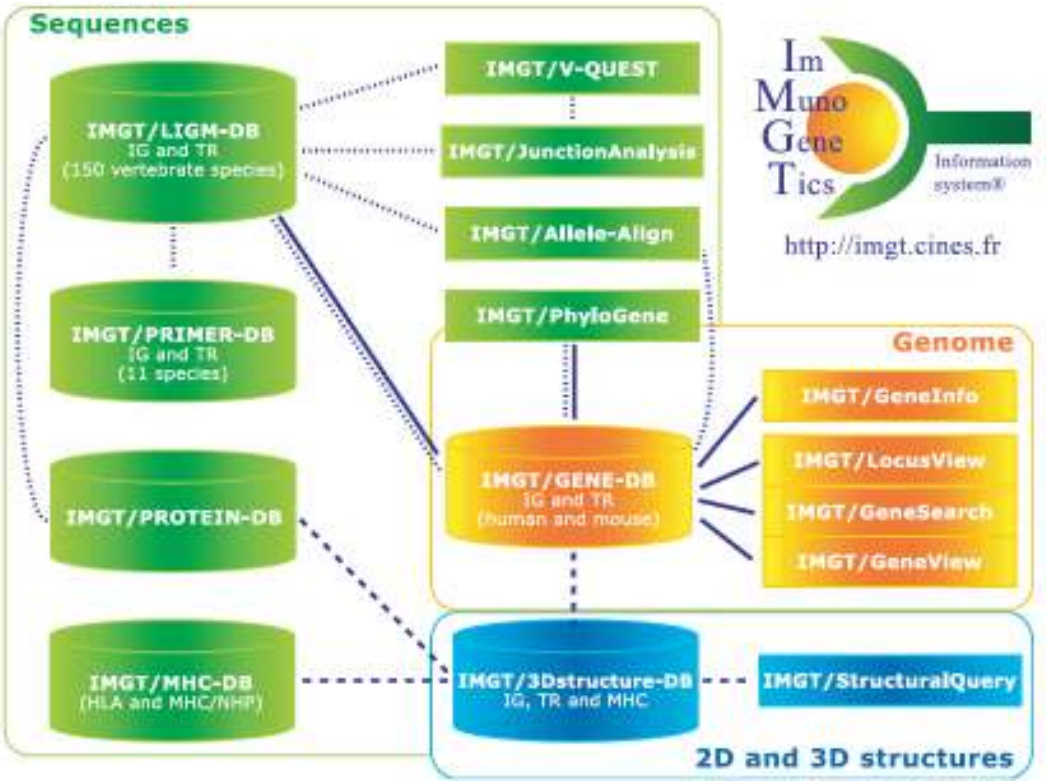


Figure 1.3: Overview of IMGT information system (Lefranc *et al*., 2005).

1.5.3 ***Immuno Polymorphism Database (IPD)***

IPD is a set of databases and tools related to immunology which are compiled to study polymorphic genes in the immune system (Robinson *et al*., 2013). IPD currently consists of four databases: IPD-KIR contains the allelic sequences of killer-cell immunoglobulin-like receptors; IPD-MHC contains sequences of the major histocompatibility complex of different species; IPD-HPA contains alloantigens expressed only on platelets; and IPD-ESTDAB provides access to the European Searchable Tumour Cell-Line Database which is a cell bank of immunologically characterized melanoma cell lines.

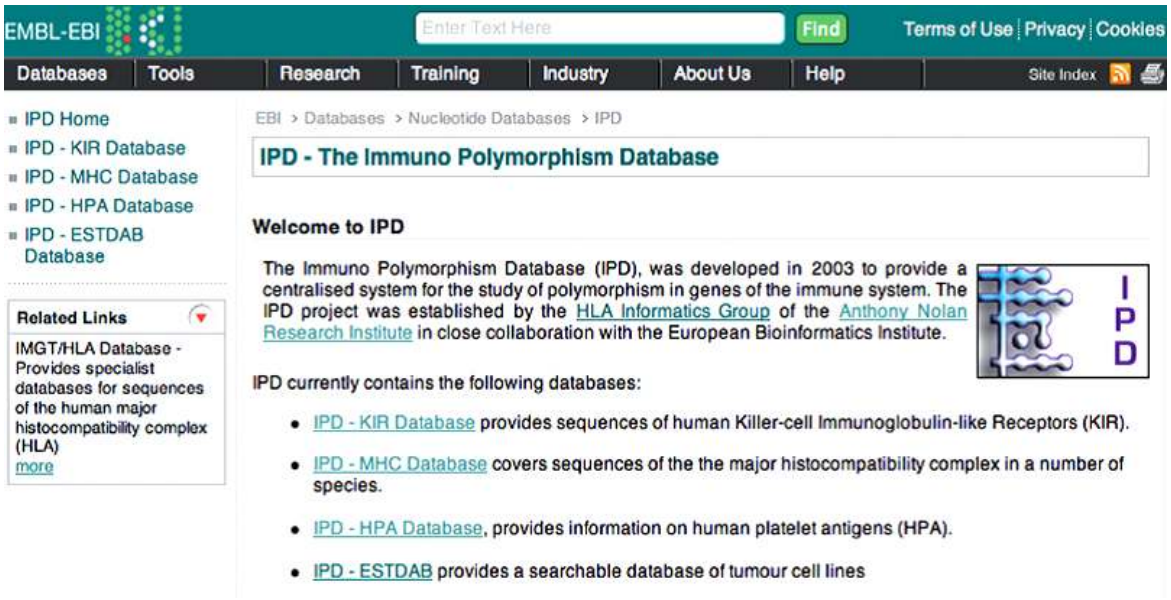


Figure 1.4: Front page of IPD database (Robinson *et al*., 2013)

Other web resources composed of mainly databases and computational tools for vaccines design are discussed in sub-sections.

1.5.4 ***Databases for vaccine design***

Databases accessible for vaccine design can be grouped under different categories such as antigen databases, epitope databases, hapten databases, MHC databases, interaction databases and miscellaneous databases. Antigen databases having information about different types of antigens such as peptide antigen (Peptide Antigen Database), polysaccharide antigen (PolysachDB), human platelet antigen (IPD-HPA), tumor related antigen (HPTAA and TANTIGEN), variation in antigens (varDB) and protective antigen (Protegen). Details of antigens databases are provided in Table 1.3. Epitope databases are categorized as under information of epitopes such as B-cell epitopes (BCIpep), all known antigenic residues and the antibodies that interact with them (Epitome), discontinuous or structural epitopes (CED or SEDB), all B-cell and T-cell epitopes (Immune Epitope Database) and 3D structure of epitopes (IEDB-3D). Details of such epitope databases are provided in Table 1.4. MHC databases contain information about MHC molecules such as sequences of the MHC from different species (IPD-MHC), sequences of the human major histocompatibility complex and nomenclature (IMGT/HLA), clinical data related to MHC (dbMHC), and MHC binding and non-binding peptides and MHCPEP (MHCBN). Details of MHC databases are provided in

Table 1.5. Hapten databases are having comprehensive information about the hapten molecule, ways to raise antibodies (HaptenDB) and 2D/3D structure of hapten (SuperHapten). Details of hapten databases are provided in Table 1.6. Interaction databases possess information related to interactions of immunological molecules such as MHC-peptide interaction database (MPID-T2), interactions and signaling pathways involved in the innate immune response (InnateDB) and the B-cell epitope interaction database (BEID). Details of interaction databases are provided in Table 1.7. Miscellaneous (other relevant) databases having information such as expression in macrophages (GPX-Macrophage Expression Atlas), X-linked severe combined immunodeficiency mutations (IL2Rgbase), information of genes (VirmugenDB) that encode for a virulent factor of a pathogen and knocking of these genes can be used to make a live attenuated vaccines, etc. Details of miscellaneous databases are provided in Table 1.8.

Table 1.3: Web resources for antigens

Name	URL	Description
AntiJen	http://www.ddg-pharmfac.net/antijen/AntiJen/antijenhomepage.htm	AntiJen v2.0, is a database containing quantitative binding data for peptides binding to MHC Ligand, TCR-MHC complexes, T cell epitope molecules, TAP, B cell epitope molecules and immunological Protein-Protein interactions. Most recently, AntiJen has included Peptide Library, Copy Numbers and Diffusion Coefficient data. All entries are from published experimentally determined data. The database currently holds over 24,000 entries (McSparron *et al*., 2003).
AntigenDB	http://www.imtech.res.in/raghava/antigendb/	Database 'AntigenDB' provides comprehensive information about a wide range of experimentally validated antigens (Ansari *et al*., 2010).

Protegen	http://www.violinet.org/protegen	This database stores experimentally validated protective antigens from archaea, prokaryotes, eukaryotes and mammals including humans. This database also contains analysis tools (Yang *et al*., 2011).
Peptide Antigen Database	http://www.proteinlounge.com/biosyn/peptide_overview.asp	Protein Lounge has created the first complete peptide-antigen database. The database has also been subdivided into folders for peptide regions for kinases, phosphatases, transcription factors and disease genes.
PolysacDB	http://crdd.osdd.net/raghava/polysacdb/	A comprehensive database of microbial polysaccharide antigens and their antibodies (Aithal *et al*., 2012).
TANTIGEN	http://cvc.dfci.harvard.edu/tadb/	Tumor T cell antigen database is a data source and analysis platform for cancer vaccine target discovery focusing on human tumor antigens that contain HLA ligands and T cell epitopes (Van den Eynde and van der Bruggen, 1997).
HPTAA	http://www.bioinfo.org.cn/hptaa/	HPTAA is a database of potential tumor-associated antigens that uses expression data from various expression platforms, including carefully chosen publicly available microarray expression data, GEO, SAGE and Unigene expression data (Wang *et al*., 2006).
varDB	http://www.vardb.org/vardb/	varDB was developed to serve as centralized database of antigenically variable protein families from a range of pathogenic organisms (Hayes *et al*., 2008).

IPD-HPA	http://www.ebi.ac.uk/ipd/hpa/	This database provides a centralized repository for the data which define the human platelet antigens (HPA). Alloantibodies against human platelet antigens are involved in neonatal alloimmune thrombocytopenia, post-transfusion purpura and refractoriness to random donor platelets (Robinson *et al*., 2005).

Table 1.4: Web resources for epitopes

Name	URL	Description
The Immune Epitope Database (IEDB)	http://www.iedb.org/	The immune epitope database (IEDB, www.iedb.org), provides a catalog of experimentally characterized B- and T-cell epitopes, as well as data on MHC binding and MHC ligand elution experiments. The database represents the molecular structures recognized by adaptive immune receptors and the experimental contexts in which these molecules were determined to be immune epitopes. Epitopes recognized in humans, non-human primates, rodents, pigs, cats and all other tested species are included. Both positive and negative experimental results are captured.
IEDB-3D	http://www.iedb.org/bb_structure.php	Structural data within the Immune Epitope Database. The B Cell, T Cell, and MHC Binding peptides are organized by the organism that is the source of the antibody, T Cell, and MHC molecule, respectively. Currently it has more than 1000 distinct molecular structures (Vita *et al*., 2010).

SEDB	http://sedb.bicpu.edu.in/	The SEDB (structural epitope database) is an open-access database for describing the three-dimensional structure of epitope containing proteins and its intermolecular contact information between antigen and antibody. It also summarizes the source information, experimental details used to determine immune responseand information on epitope such as sequence and visualization of epitope (Premendu, 2012).
Bcipep	http://bioinformatics.uams.edu/mirror/bcipep/	A database of immunodominant B cell epitopes (Saha *et al*., 2005).
Epitome	https://rostlab.org/services/epitome/	This is a database of all known antigenic residues and the antibodies that interact with those residues, including a detailed description of residues involved in the interactions and their sequence / structure environments. Additionally, interactions can be visualized using a visualization interface, Jmol.
CED	http://immunet.cn/ced/	Conformational epitope database (CED) provides a collection of conformational epitopes producing antibody immune response and related information including the immunological property of the epitope, the residue make up and location of the epitope, and the source antigen and corresponding antibody of the epitope (Huang and Honda, 2006).

Table 1.5: Web resources for major histocompatibility complex (MHC)

Name	URL	Description
dbMHC	http://www.ncbi.nlm.nih.gov/gv/mhc/main.fcgi?cmd=init	This database provides an open and publicly accessible platform for DNA and clinical data related to the human major histocompatibility complex (MHC).
MHCBN	http://www.imtech.res.in/raghava/mhcbn/	This is a curated database consisting of detailed information about major histocompatibility complex (MHC) binding, non-binding peptides and T-cell epitopes. The version 4.0 of database also provides information about peptides interacting with TAP and possibility of peptide binding (Bhasin *et al*., 2003).
MHCPEP	http://wehih.wehi.edu.au/mhcpep/	It is a curated database comprising over 4000 peptide sequences known to bind MHC molecules. Entries are compiled from published reports as well as from direct submissions of experimental data (Brusic *et al*., 1998).
IMGT/HLA	http://www.ebi.ac.uk/ipd/imgt/hla/	This is a specialist database for providing sequences of the human major histocompatibility complex (hMHC) and includes the official sequences for the WHO Nomenclature Committee for Factors of the HLA System. This database is part of the international ImMunoGeneTics project (IMGT) (Robinson *et al*., 2003).
IPD-MHC	http://www.ebi.ac.uk/ipd/mhc/	This database provides a centralized repository for sequences of the major histocompatibility complex (MHC) from a number of different species. Through a

		number of international collaborations, the IPD is able to provide the MHC sequences from different species. The sequences provided by each group are curated by experts in the field and then submitted to the central database (Robinson *et al*., 2005).

Table 1.6: Web resources for hapten

Name	URL	Description
SuperHapten	http://bioinformatics.charite.de/superhapten/	SuperHapten is a manually curated hapten database integrating information from literature and web resources. The current version of the database compiles 2D/3D structures, physicochemical properties and references for about 7,500 haptens and 25,000 synonyms (Günther *et al*., 2007).
HaptenDB	http://www.imtech.res.in/raghava/haptendb/	This is a database of haptens which provides comprehensive information about the hapten molecule (ways to raise antibodies against particular group of haptens), specificity and cross reactivity of raised antibody with related haptens) (Singh *et al*., 2006).

Table 1.7: Web resources for Interactions in immunological molecules

Name	URL	Description
InnateDB	http://www.innatedb.com/	InnateDB is a publicly available database of the genes, proteins, experimentally-verified interactions and signaling pathways involved in the innate immune response of humans, mice and bovines to microbial infection (Lynn *et al*., 2008).
MPID-T2	http://biolinfo.org/mpid-t2/	The MHC-Peptide Interaction Database-TR

		version 2 (MPID-T2) is a new generation database for sequence-structure-function information on T cell receptor/peptide/MHC interactions. It contains all known crystal structures of TR/pMHC and pMHC complexes with emphasis on the structural characterization of these complexes (Khan *et al*., 2011).
The B-Cell Epitope Interaction Database (BEID)	http://datam.i2r.a-star.edu.sg/BEID/	This is a new generation database for structure-function information on B-cell epitope interactions. It contains structures of immunoglobulin (Ig)-antigen complexes with emphasis on the structural characterization of these complexes (Tong *et al*., 2008).

Table 1.8: Miscellaneous databases in immunology

Name	URL	Description
IL2Rgbase	http://www.ncbi.nlm.nih.gov/lovd/home.php?select_db=IL2RG	A database of human X-SCID mutations (IL2RGbase) has been assembled, and this article summarizes the first 136 entries from unrelated patients (Puck *et al*., 1996).
VBASE2	http://www.vbase2.org/	It is an integrative database of germ-line V genes from the immunoglobulin loci of human and mouse. It presents V gene sequences from both EMBL database and Ensembl with the corresponding links to the source data (Retter *et al*., 2005).
GPX-Macrophage Expression Atlas	http://gpxmea.gti.ed.ac.uk/	This macrophage expression atlas (database) provides expression profiles of macrophages challenged with a variety of pro-inflammatory, anti-inflammatory, benign and pathogen insults (Grimes *et al*., 2005).

DIGIT	http://www.biocomputing.it/digit4/ (not working)	Database of ImmunoGlobulin sequences and Integrated Tools. It is an integrated resource storing sequences of annotated immunoglobulin variable domains and enriched with tools for searching and analyzing them. (Chailyan *et al*., 2012).
MUGEN Mouse Database (MMdb)	http://www.mugen-noe.org/database/ (not working)	Murine models of immune processes and immunological diseases. MMdb's basic classification of models is based on three major research application categories: Models of Human Disease, Models of Immune Processes and Transgenic Tools. (Aidinis *et al*., 2008).
VirmugenDB	http://www.violinet.org/virmugendb/index.php4	"Virmugen" is coined here to represent a gene that encodes for a virulent factor of a pathogen. It has already been proven regarding feasibility of making a live attenuated vaccine by knocking out this gene. Not all virulence factors can be used for vaccine development. Currently, VIOLIN includes 225 Virmugens that were verified to be valuable for vaccine development against 57 pathogens (Racz *et al*., 2012).

1.5.5 ***Computational tools for vaccine design***

The computational tools (implemented as web servers or standalone softwares) are required for analysis and predictions in immunological studies which is generally useful in vaccine design. Especially, computational tools used for vaccine candidate predictions are instrumental for design of efficient bench experiments which sometime assists to develop vaccine design pipelines (Wee *et al*., 2012). These tools are not only reducing time and cost associated with research experiments, but these computational methods are also helping to cross the obvious limitations of experimental research. According to applications, the immunoinformatics tools can be clustered into different categories such as antigen prediction, allergen prediction, T-cell epitopes prediction and B-cell epitopes predictions.

1.5.5.1 Antigen prediction tools

Antigen prediction tools are ANTIGENpro, VaxiJen, new enhanced reverse vaccinology environment (NERVE) and Vaxign. ANTIGENpro predicts antigenic proteins while NERVE and VaxiJen predict protective antigenic proteins. Both VaxiJen and ANTIGENpro were based on machine leaning techniques: VaxiJen uses alignment independent method; discriminant analysis and partial least square (DA-PLS), for prediction of antigen (Doytchinova and Flower, 2007) whereas ANTIGENpro uses support vector machine classifier for prediction (Magnan *et al*., 2010). NERVE and Vaxign use reverse vaccinology (RV) principle for the prediction (He *et al*., 2010; Vivona *et al*., 2006). Details of antigen prediction tools are provided in Table 1.9.

Table 1.9: Tools for antigenic protein prediction

Name	URL	Description
ANTIGENpro	http://scratch.proteomics.ics.uci.edu/	This tool Predicts antigenicity through machine learning technique (Magnan *et al*., 2010).
NERVE	http://www.bio.unipd.it/molbinfo/	This software predicts protein vaccine candidates through localization, number of transmembrane helices and adhesion likeliness (Vivona *et al*., 2006).
VaxiJen	http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html	This web server predicts protective antigens and subunit vaccine candidates using discriminant analysis and partial least square (DA-PLS) methods (Doytchinova and Flower, 2007).
Vaxign	http://www.violinet.org/vaxign/	This web resource is a vaccine target prediction and analysis system based on the principle of reverse vaccinology. Two programs exist in Vaxign: 1. 'Vaxign Query' provides pre-computed Vaxign results for user to explore and 2. 'Dynamic Vaxign Analysis'

		can take input sequences from user and perform dynamic Vaxign execution, and have provision for result visualization (He *et al*., 2010).

1.5.5.2 Allergen prediction tools

An allergen is a type of antigen which can produces an unusual vigorous immune response. Allergen stimulate type-I hypersensitivity reaction through Immunoglobulin E (IgE) responses and can causes allergy. World Health Organization (WHO) and the Food and Agriculture Organization (FAO) proposed guidelines to assess the potential allergencity of proteins (Saha and Raghava, 2006a).

Allergen prediction tools are AllerHunter, AlgPred and EVALLER. The AllerHunter is available as web based application, and it predicts allergenicity and allergic cross-reactivity of proteins through SVM classifier (Muh *et al*., 2009). AlgPred uses different approaches such as presence of IgE epitopes, similarity with allergen peptides, motif search and machine learning technique for allergen prediction (Saha and Raghava, 2006a). EVALLER incorporates filtered length-adjusted allergen peptides (DFLAP) algorithm for allergen prediction (Barrio *et al*., 2007). Details of tools for allergen prediction are provided in Table 1.10.

Table 1.10: Tools for allergen prediction

Name	URL	Description
AllerHunter	http://tiger.dbs.nus.edu.sg/AllerHunter/	This web server predicts allergenicity and allergic cross-reactivity of proteins. It combines an iterative pairwise sequence similarity encoding scheme with SVM as the classification engine (Muh *et al*., 2009).
AlgPred	http://www.imtech.res.in/raghava/algpred/	This tool uses combination of approaches such as presence of IgE epitopes, similarity with allergen peptides, motif search and machine learning based classification for allergen prediction (Saha and Raghava, 2006a).

EVALLER	http://bioinformatics.bmc.uu.se/evaller.html	This web server uses filtered length-adjusted allergen peptides (DFLAP) algorithm for allergen prediction (Barrio *et al*., 2007).

1.5.5.3 Discontinuous B-cell epitopes prediction tools

B-cell epitopes (BCEs) predictions are grouped into linear and discontinuous or structural epitopes prediction methods. Majority (90%) of the reported BCEs is discontinuous epitopes. Therefore, the most of the developed BCE prediction tools are discontinuous or structural B-cell epitopes, and the name of such prediction tools are Ellipro, BEpro, DiscoTope 2.0, SEPPA, EPSVR, EPMeta, Bpredictor, EPITORIA, EPCES and CBTOPE. These tools generally require three dimensional structure of protein usually in PDB format as input. These tools are based on different methods. Machine learning techniques, like naive Bayes classifier (tool EPITORIA), random forests with a distance-based feature (Zhang *et al*., 2011), and support vector machine classification (CBTOPE) (Ansari and Raghava, 2010) and SVM regression (EPSVR) (Liang *et al*., 2010) are used for prediction. Surface accessibility (estimated in terms of contact numbers) and a novel epitope propensity amino acid score used in DiscoTope 2.0 (Kringelum *et al*., 2012)., Thornton's method with a residue clustering algorithm used in Ellipro (Ponomarenko *et al*., 2008), Concept of 'unit patch of residue triangle' to describe the local spatial context in protein surface and clustering coefficient to describe the spatial compactness of surface residues were used in SEPPA (Sun *et al*., 2009) are implemented for predictions. Details of the tools for discontinuous B-cell epitopes prediction are provided in Table 1.11.

Table 1.11: Tools for discontinuous or conformational B-cell epitope prediction

Name	URL	Description
Ellipro	http://tools.immuneepitope.org/tools/ElliPro/iedb_input	This tool predicts linear and discontinuous antibody epitopes based on a protein antigen's 3D structure. It associates each predicted epitope with a score, defined as a PI (Protrusion Index) value averaged over epitope residues (Ponomarenko *et al*., 2008).
BEpro	http://pepito.proteomic	This web server (formerly known as PEPITO) is a predictor of discontinuous B-cell epitopes (BCEs). All

	s.ics.uci.edu/	that is required is the tertiary structure of the antigen in the PDB format (Sweredoski and Baldi, 2008).
CPC-BCE	http://bcell.whu.edu.cn/sequence_input.html	Computational prediction of conformational B-cell epitopes from antigen primary structures by using ensemble learning (Zhang *et al*., 2012b)
DiscoTope 2.0	http://www.cbs.dtu.dk/services/DiscoTope/	This web server predicts discontinuous BCEs from protein three dimensional structures. The method utilizes calculation of surface accessibility (estimated in terms of contact numbers) and a novel epitope propensity amino acid score. The final scores are calculated by combining the propensity scores of residues in spatial proximity and the contact numbers (Kringelum *et al*., 2012).
SEPPA	http://lifecenter.sgst.cn/seppa/	The SEPPA (Spatial Epitope Prediction of Protein Antigens) server is used as a tool for conformational B-cell epitope prediction. With 3D protein structure as input, each residue in the query protein will be given a score according to its neighborhood residues' information. Higher score corresponds to higher probability of the residue to be part of an epitope (Sun *et al*., 2009).
EPSVR	http://sysbio.unl.edu/EPSVR/	Discontinuous antigenic epitopes prediction using support vector regression (Liang *et al*., 2010)
EPMeta	http://sysbio.unl.edu/EPMeta/	A meta server for prediction of discontinuous antigenic epitopes. It combined EPSVR with five existing epitope prediction servers to construct EPMeta (Liang *et al*., 2010).
Bpredictor	http://code.google.com/p/my-project-bpredictor/downloads/list	A new method to identify the B-cell conformational epitopes from 3D structures by combining conventional and the proposed features. The random forest (RF) algorithm is used as the classification engine (Zhang *et al*., 2011).
Epitopia	http://epitopia.tau.ac.il	This server implements a machine-learning algorithm

		which was trained to discern antigenic features within a given protein. A special emphasis was put on the development of a user-friendly graphical interface for displaying the results (Rubinstein *et al*., 2009).
EPCES	http://sysbio.unl.edu/EPCES/	Prediction of antigenic epitopes on protein surfaces by consensus scoring approach (Liang *et al*., 2009).
CBTOPE	http://www.imtech.res.in/raghava/cbtope/	First method which can predict conformational B cell epitope without using structure or structure of homolog. It uses amino acid composition for SVM based prediction (Ansari and Raghava, 2010).

1.5.5.4 Continuous B-cell epitope prediction tools

Continuous or linear B-cell epitope predictions tools are Antibody Epitope Prediction, BCPREDS, BcePred, ABCpred, LBtope, BepiPred 1.0 Server, BayesB (version 1.0), COBEpro, and BEST. These prediction tools were based on different techniques: Artificial neural network (ANN) is used in ABCpred server (Saha and Raghava, 2006b); "Antibody Epitope Prediction" server in IEDB provides option to use different methods (Chou & Fasman beta-turn prediction, Emini surface accessibility prediction, Karplus & Schulz flexibility prediction, Kolaskar & Tongaonkar antigenicity, Parker hydrophilicity prediction and Bepipred linear epitope prediction); amino acid pair (AAP) immunogenicity scale is used in Bcepred (Saha and Raghava, 2004); BayesB methods uses Bayes feature extraction for encoding the feature vectors in the support vector machines algorithm to predict linear epitopes (Wee *et al*., 2010); COBEpro uses support vector machine initially and then calculates an epitopic propensity score for each residue based on the fragment predictions (Sweredoski and Baldi, 2009); BepiPred predicts linear B-cell epitopes using a combination of a hidden Markov model and a propensity scale method (Larsen *et al*., 2006); and BEST method is based on support vector machine (Gao *et al*., 2012). Details of tools for continuous or linear B-cell epitopes prediction are provided in Table 1.12.

Table 1.12: Tools for linear B-cell epitope prediction

Name	URL	Description
Antibody Epitope Prediction	http://tools.immuneepitope.org/tools/bcell/iedb_input	The following methods are provided for B-cell epitope predictions: 1. Chou & Fasman beta-turn prediction 2. Emini surface accessibility prediction 3. Karplus & Schulz flexibility prediction 4. Kolaskar & Tongaonkar antigenicity 5. Parker hydrophilicity prediction 6. Bepipred linear epitope prediction
BCPREDS	http://ailab.cs.iastate.edu/bcpreds/	BCPREDS (B-cell epitope prediction server) allows users to choose the method for predicting B-cell epitopes. The current implementation of BCPREDS allows the user to select among three prediction methods including our implementation of amino acid pair (AAP) method (EL-Manzalawy *et al*., 2008).
Bcepred	www.imtech.res.in/raghava/bcepred	The aim of this server is to predict B cell epitope regions in an antigen sequence, using physico-chemical properties. This server has been tested on B cell epitope database, BCIPEP (www.imtech.res.in/raghava/bcipep). It can also predict continuous B cell epitopes. Identified properties of B cell epitope include hydrophilicity, flexibility/mobility, accessibility, polarity, exposed surface and turns. Quantification of these properties is determined by assigning a value to each of the 20 natural amino acids (Saha and Raghava, 2004).
ABCpred	http://www.imtech.res.in/raghava/abcpred/	The aim of ABCpred server is to predict B cell epitope(s) in an antigen sequence, using

		artificial neural network. This is the first server developed based on recurrent neural network (machine learning technique) using fixed length patterns (Saha and Raghava, 2006b).
LBtope	http://crdd.osdd.net/raghava/lbtope/	This web server (LBtop) is used for predicting linear B-cell epitopes and experimentally validated non B-cell epitopes were used first time for developing prediction model. Several models were developed using various techniques (*e.g.* SVM, IBk) on a large dataset of B-cell epitopes (12063) and non-epitopes (20589) from IEDB database (Singh *et al*., 2013).
BepiPred 1.0	http://www.cbs.dtu.dk/services/BepiPred/	This server predicts the location of linear B-cell epitopes using a combination of a hidden Markov model and a propensity scale method (Larsen *et al*., 2006).
BayesB (V 1.0)	http://immunopred.org/bayesb/index.html	This web server predicts linear B-cell epitopes on protein sequence. It employs the use of Bayes feature extraction for encoding the feature vectors in the support vector machines algorithm. The best prediction model was attained on 20-mer window (Wee *et al*., 2010).
COBEpro	http://scratch.proteomics.ics.uci.edu/	It is a novel two-step system for predicting continuous B-cell epitopes. Epitopic propensity scores are assigned to both standalone peptide fragments and residues within an antigen sequence. The support vector machine is used first to make predictions on short peptide fragments within the query antigen sequence and then epitopic propensity score is calculated for each residue based on the fragment

		predictions (Sweredoski and Baldi, 2009).
BEST	http://biomine.ece.ualberta.ca/BEST/	This method combines information derived from the chain, sequence conservation, similarity to known (training) epitopes, and predicted secondary structure and relative solvent accessibility to predict B-Cell epitopes from antigen sequences (BEST) (Gao *et al*., 2012).

1.5.5.5 T-cell epitope prediction tools

T-cell epitopes prediction can be divided into MHC class I and MHC class II types and prediction accuracy of the former is better than the latter. Currently, different methods and algorithms are available for T-cell epitope predictions. The MHC binding prediction methods are mainly based on support vector machine (SVM). MHC2Pred is a SVM method to predict HLA-DRB1(*)0401 binding peptides in an antigen sequence (Bhasin and Raghava, 2004b). MHC class I binding and proteosomal cleavage is performed using artificial neural networks (ANNs), and TAP transport efficiency is predicted using weight matrix in NetCTL (Larsen *et al*., 2007). NetMHC, NetMHCpan and NetMHCII use the ANNs to predict binding of peptides with number of different HLA alleles (Hoof *et al*., 2009; Larsen *et al*., 2006; Lundegaard *et al*., 2008; Nielsen and Lund, 2009). NetMHCIIpan uses pan-specific method capable of predicting peptide binding to any HLA class II molecule with a defined protein sequence (Karosiene *et al*., 2013). SVMHC has MHC class I prediction which are based on support vector machines (SVMs) and known MHC-binding peptide (Dönnes and Elofsson, 2002). SVRMHC utilizes quantitative method of modeling, the interaction between a peptide and a MHC molecule, based on the support vector machine regression (SVR) method (Wan *et al*., 2006). Physicochemical properties from known MHC class I binding peptides were used to design a support vector machine (SVM) based system in POPI (Tung and Ho, 2007). CTLpred method is based on elegant machine learning techniques like ANN and SVM (Bhasin and Raghava, 2004a). A neural network based MHC Class I binding peptide prediction method used in nHLAPred (Bhasin and Raghava, 2007). Linear programming method is employed for predicting HLA-A2 binding peptides in SMM. (Peters *et al*., 2003). Position specific scoring matrices (PSSMs) and proteosomal cleavage is used in RANKPEP (Reche *et al*., 2002) method. ProPred incorporates matrix based prediction algorithm, using

amino-acid/position coefficient table deduced from literature, for prediction of MHC Class I epitopes (Singh and Raghava, 2003). Details of tools for T-cell epitopes prediction are provided in Table 1.13.

Table 1.13: Tools for T-cell epitope prediction

Name	URL	Description
NetMHC 3.4 Server	http://www.cbs.dtu.dk/services/NetMHC/	This web server is based on accurate approximation method for prediction of Class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers (Lundegaard *et al*., 2008).
RANKPEP	http://bio.dfci.harvard.edu/RANKPEP/	This server predicts MHC I and MHC II peptide binders from protein sequence or sequence alignments using position specific scoring matrices (PSSMs). In addition, it predicts those MHC I ligands with a C-terminal end that is likely to be the result of proteasomal cleavage (Reche *et al*., 2002).
NetCTL-1.2	http://www.cbs.dtu.dk/services/NetCTL/	This server uses method that integrates prediction of peptide MHC class I binding, proteasomal C terminal cleavage and TAP transport efficiency. MHC class I binding and proteasomal cleavage is performed using artificial neural networks whereas TAP transport efficiency is predicted using weight matrix (Larsen *et al*., 2007).
MHC2Pred	http://www.imtech.res.in/raghava/mhc2pred/	SVM based method for predicting HLA-DRB1(*)0401 binding peptides in an antigen sequence (Bhasin and Raghava, 2004b).
PREDNOD	http://antigen.i2r.a-star.edu.sg/Ag7	A prediction server for peptide binding to the H-2g7 haplotype of the non-obese diabetic mouse (Rajapakse *et al*., 2006).
SYFPEITHI	http://www.syfpeithi.de/bin/MHCServer.dll/EpitopePrediction.htm	This web server predicts T-cell epitopes based on probability of being processed and presented (Rammensee *et al*., 1999).

NetMHCII 2.2 Server	http://www.cbs.dtu.dk/services/NetMHCII/	This web server predicts binding of peptides to HLA-DR, HLA-DQ, HLA-DP and mouse MHC class II alleles using artificial neural networks (Nielsen and Lund, 2009).
NetMHCpan 2.8 Server	http://www.cbs.dtu.dk/services/NetMHCpan/	This web server predicts binding of peptides to any known MHC molecule using ANNs. The method is trained on more than 150,000 quantitative binding data covering more than 150 different MHC molecules. Predictions can be made for HLA-A, B, C, E and G alleles, as well as for non-human primates, mouse, cattle and pig (Hoof *et al*., 2009).
NetMHCIIpan 3.0 Server	http://www.cbs.dtu.dk/services/NetMHCIIpan/	The NetMHCIIpan 3.0 server predicts binding of peptides to MHC Class II molecules. The predictions are available for all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ, as well as mouse molecules (Karosiene *et al*., 2013).
ProPred	http://www.imtech.res.in/raghava/propred/	The aim of this server is to predict MHC Class II binding regions in an antigen sequence using quantitative matrices (Singh and Raghava, 2001).
ProPred-I	www.imtech.res.in/raghava/propred1/	This web server is used for identifying the MHC Class I binding regions in antigens. It implements matrices for 47 MHC Class I alleles, proteosomal and immuneproteosomal models (Singh and Raghava, 2003).
SVMHC	http://abi.inf.uni-tuebingen.de/Services/SVMHC/information	This web server is capable of predicting both MHC class I and MHC class II binding peptides. The graphical output also allows for simple identification of promiscuous epitopes (Dönnes and Elofsson, 2002).
SVRMHC	http://svrmhc.biolead.org/	This web server predicts peptide-MHC binding affinities using SVRMHC models. Currently, 36

		class I SVRMHC models and 6 class II SVRMHC models are hosted here (Wan *et al*., 2006).
POPI2.0	http://iclab.life.nctu.edu.tw/POPI/	A web server for predicting immunogenicity of MHC class I and II binding peptides through mining of informative physicochemical properties (Tung and Ho, 2007).
CTLpred	http://www.imtech.res.in/raghava/ctlpred/	This server uses two methods for prediction. In direct method, the information or patterns of T cell epitopes instead of MHC binders were used whereas the consensus method combined the prediction of both ANNs and SVM for the prediction (Bhasin and Raghava, 2004a).
nHLAPred	http://www.imtech.res.in/raghava/nhlapred/	A neural network server based on MHC Class I binding peptide prediction (Bhasin and Raghava, 2007).
SMM	http://cagt.bu.edu/page/SMM_submit	This server uses linear programming method for predicting HLA-A2 binding peptides (Peters *et al*., 2003).

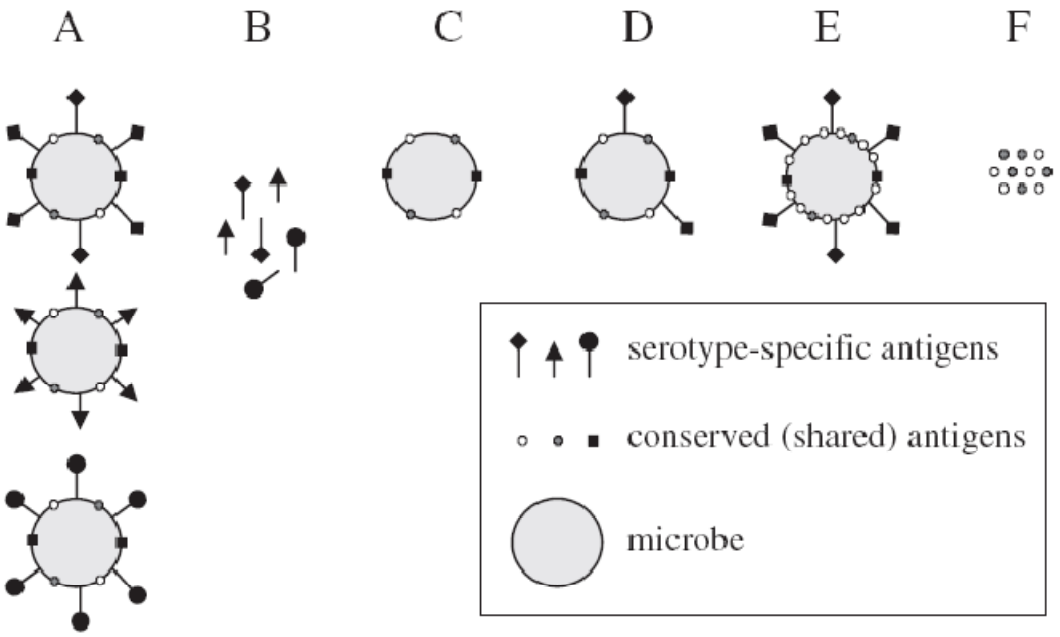
1.6 Broad specific vaccine design

Conventional vaccines based on outer surface antigens are generally sero-dependent as sero-specific polysaccharide antigens are used for formulation of such vaccines. In such cases, serotypes not covered by the vaccines become more frequently colonize due to selective pressure and turn out to be an important factor for infection. For example, conjugate vaccines for *Neisseria meningitidis* were administered against serogroup C and B in Europe and the United States respectively. But a decade later, serogroup Y became prevalent in the United States and vaccines containing both C- and Y-serotypes can only provide protective immunity in the US. As pathogens are evolving to novel serotypes and strains, broad-specific vaccines are essential for getting protection against particular organisms (Rappuoli, 2007). Earlier there was a general perceived notion that sero-specific outer surface proteins only provide protective immunity. In post-genomics era, immunomics revealed that protective immunity is not only confined to virulence determinant. Antigenic elements may involve

range of other different proteins such as housekeeping, structural and functional proteins, and even proteins with unknown function (De Groot and Moise, 2007). Housekeeping proteins are conserved and can be mined through pan genomic approaches. Any vaccine candidates identified from such conserved class of proteins can provide protection against a range of pathogens or strains.

Several strategies have been proposed to design broad specific vaccines in literature (Figure 1.5). Polyvalent vaccine contained the combination of all different serotypes (Figure 1.5 (**A**)) or purified serotype-determining antigens from a given pathogen or set of pathogens (**B**). This type of vaccine development is feasible in case of those pathogens which have few serotypes/variants *i.e. Haemophilus influenzae* and polio vaccine (Nagy and Pál 2008). But in case of highly variable viruses (such as influenza A virus and HIV) which have several serotypes or subtypes, the practical feasibility is not yet possible through polyvalent vaccines, and highly conserved as well as immunogenic antigens are required for broad-specific vaccine development. Abolishment (Figure 1.5 (**C**)) or down-regulation (Figure 1.5 (**D**)) of sero-specific immunogenic antigens generally results in increase of the immunogenicity of low immunogenic but conserved antigen. In general, the over-expression of conserved antigen also provides broad specific immune responses (Figure 1.5 (**E**)). Over-expression of protective antigen can be used to improve the efficacy of vaccines (Vemulapalli *et al*., 2000). Finally use of purified conserved antigens as subunit vaccine has potential to be used as broad protective vaccines (Figure 1.5 (**F**)).

In several pathogens, these purified conserved antigens based vaccines are being explored for vaccines development: Porins are explored as broad spectrum vaccines against Salmonella and outer membrane proteins (OMPs) are explored vaccine candidates against several pathogens including pathogenic Leptospira and *Haemophilus influenza* (El-Adhami *et al*., 1999; Gebriel *et al*., 2006). Over-expression of protective antigen has been used as new way to modulate immune response to increase efficacy of Brucella and *Mycobacterium tuberculosis* vaccines (Rao *et al*., 2003; Vemulapalli *et al*., 2000). In highly variable viruses, like influenza, highly conserved among strains as well as immunogenic epitopes were combined to develop a broad-specific vaccine, Multimeric-001, which is in Phase-II clinical trial . Any of the proposed method for broad-spectrum vaccine development such as polyvalent vaccine or conserved antigenic vaccine or any new approach against highly variable pathogens including viruses can be possible through computational support.



(A) Multiple serovariants of a given organism are combined into polyvalent vaccines.

(B) Purified serotype-specific antigens thereof are combined into polyvalent vaccines.

(C) Expression of the multiform surface antigens responsible for serotype-variability can be abolished.

(D) Expression of the multiform surface antigens responsible for serotype-variability can be down-regulated.

(E) Conserved antigens may be over-expressed in an attenuated parental strain or a suitable heterologous vector strain.

(F) Purified conserved antigens shared by all serotypes could be used as subunit vaccines.

Figure 1.5: Strategies to design broad protective vaccines (Nagy and Pál, 2008).

1.7 Limitation of current Immunoinformatics tools

Though the reverse vaccinology (RV) method was instrumental in providing better vaccine candidates against dozens of human and animal pathogens including *Neisseria meningitidis* (Pizza *et al*., 2000), *Helicobacter pylori* (Chakravarti *et al*., 2000), *Streptococcus pneumoniae* (Wizemann *et al*., 2001), *Porphyromonas gingivalis* (Ross *et al*., 2001), *Chlamydia pneumoniae* (Montigiani *et al*., 2002), *Bacillus anthracis* (Ariel *et al*., 2002), *Leptospira interrogans* (Yang *et al*., 2006), etc. but this method needs huge number of experiments which are time consuming and costly for identification of vaccine candidates.

Current tools based on RV did not provide explicit analysis on possible immunogenic potential in terms of immunogenic region(s) (Doytchinova and Flower, 2007; He *et al*., 2010; Vivona *et al*., 2006). This method also did not provide conservancy of vaccine candidates across different strains in terms of pathogenic and non-pathogenic. Therefore, the RV techniques focused to develop better antigen prediction and analyses of antigens are required.

The role of computational methods has become a decisive factor in vaccine design. Dependency on computational tools demands high accuracy and applicability to encounter challenging diseases through vaccines. Current antigen prediction methods such as ANTIGENpro, NERVE, VaxiJen and Vaxign are either based on machine learning techniques or prediction of adhesion likeliness for identifying antigen (Doytchinova and Flower, 2007; He *et al*., 2010; Magnan *et al*., 2010; Vivona *et al*., 2006). These methods do not consider all biological aspect for antigen prediction. Accuracy of antigen prediction can be enhanced through considering different biological aspect such host pathogen interaction and pathogenesis.

Initially epitope prediction was considered as important factor for vaccine design research but low prediction accuracy limits the application of these methods. Over-prediction, inability in exact position prediction of epitopes and absence of success in identifying known epitopes in proteins are major concerns in vaccine candidate identification. B-cell epitopes are mainly discontinuous epitopes and computational tools used for prediction discontinuous B-cell epitopes are not accurate (Blythe and Flower, 2005; Zhang *et al*., 2011). T-cell epitopes prediction tools mainly MHC II binding prediction are also suffering from low prediction accuracy (Gowthaman and Agrewala, 2007). Currently, there are many computational resources *i.e.* IEDB, IGMT, IPD, etc. which are maintaining high throughput data on experimentally known immunogenic epitopes and other information. Instead of using epitope prediction methods such types data can be used for identifying immunogenic regions in an antigen.

In viruses, antigen and epitope predictions are not the major concerns as they have very few numbers if proteins. But lack of effectiveness of vaccines against viruses is mainly due to antigenic variability which arises mostly due to antigenic shifts and drifts. There is no computational tool which directly provides conservation of antigens or epitopes to directly address antigenic variability. The conservation data of epitope is crucial for vaccine design for highly variable viruses and bacteria. Multimeric-001 developed by BiondVax Company, vaccine against influenza which is currently in clinical trials, is a vaccine based on conserved

epitopes has again heaved the opportunity of conserved epitopes based vaccine as universal influenza vaccines (Atsmon *et al*., 2012).

These limitations associated with current computational tools, and availability of validated web resources on antigenic elements and their experimental data demands the development of new and accurate prediction methods, and web resources based on biological aspects.

OBJECTIVES:

1. To develop a new method for bacterial protein vaccine candidate (PVC) prediction using information critical to host-pathogen interaction and/or pathogenesis.
2. To develop a knowledge-based resource for influenza virus and combining it with new algorithm (forward selection algorithm (FSA)) to assist design of universal epitope based vaccine against influenza.
3. Design and Implementation of web resources of the above new methods to create cybertools.

OUTLINE OF THESIS

The above objectives were successfully accomplished by developing appropriate methodologies, and implementing those methods to develop web resources and servers which predict protein vaccine candidates Jenner-Predict (http://14.139.240.55/vaccine/home.html) discussed in Chapter 2 and design of potential universal influenza vaccine candidates through EpiCombFlu web resource (http://14.139.240.55/influenza/home.html) discussed in Chapter 3.

CHAPTER 2

JENNER-PREDICT SERVER: PREDICTION OF PROTEIN VACCINE CANDIDATES (PVCS) IN BACTERIA BASED ON HOST-PATHOGEN INTERACTIONS

ABSTRACT

Subunit vaccines based on recombinant proteins have been effective in preventing infectious diseases and are expected to meet the demands of future vaccine development. Computational approach, especially reverse vaccinology (RV) method has enormous potential for identification of protein vaccine candidates (PVCs) from a proteome of a pathogenic organism. The existing protective antigen prediction software and web servers have low prediction accuracy leading to limited applications in vaccine development. Besides machine learning techniques, existing softwares and web servers have considered only protein's adhesin-likeliness as criterion for identification of PVCs. Several non-adhesin functional classes of proteins involved in host-pathogen interactions and pathogenesis are known to provide protection against bacterial infections. Therefore, knowledge of bacterial pathogenesis has potential to identify PVCs.

A web server, Jenner-Predict, has been developed for prediction of PVCs from proteomes of bacterial pathogens. The web server targets host-pathogen interactions and pathogenesis by considering known functional domains from protein functional classes such as adhesin, virulence, invasin, porin, flagellin, colonization, toxin, choline-binding, penicillin-binding, transferring-binding, fibronectin-binding and solute-binding. It predicts non-cytosolic proteins containing above mentioned domains as PVCs. It also provides vaccine potential of predicted PVCs in terms of their possible immunogenicity by comparing with experimentally known IEDB epitopes, absence of autoimmunity and their conservation in different strains. Predicted PVCs are prioritized so that only few prospective PVCs could be validated experimentally.

The performance of web server was evaluated against known protective antigens from diverse classes of bacteria reported in Protegen database and datasets used for VaxiJen server development. The web server efficiently predicted known vaccine candidates reported from *Streptococcus pneumoniae* and *Escherichia coli* proteomes. The Jenner-Predict server has outperformed other comparative (NERVE, Vaxign and VaxiJen) methods. It has sensitivity of 0.774 and 0.711 for Protegen and VaxiJen dataset, respectively while specificity of 0.940 has been obtained for the latter dataset. Better prediction accuracy of Jenner-Predict web server signifies that domains involved in host-pathogen interactions and pathogenesis are better criteria for prediction of PVCs. The web server has successfully predicted maximum known PVCs belonging to different functional classes. Jenner-Predict server is freely accessible at http://14.139.240.55/vaccine/home.html.

(The content of this chapter has already been published: V. Jaiswal, *et al*., (2013). **"Jenner-Predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions."** *BMC Bioinformatics* **14**: 211)

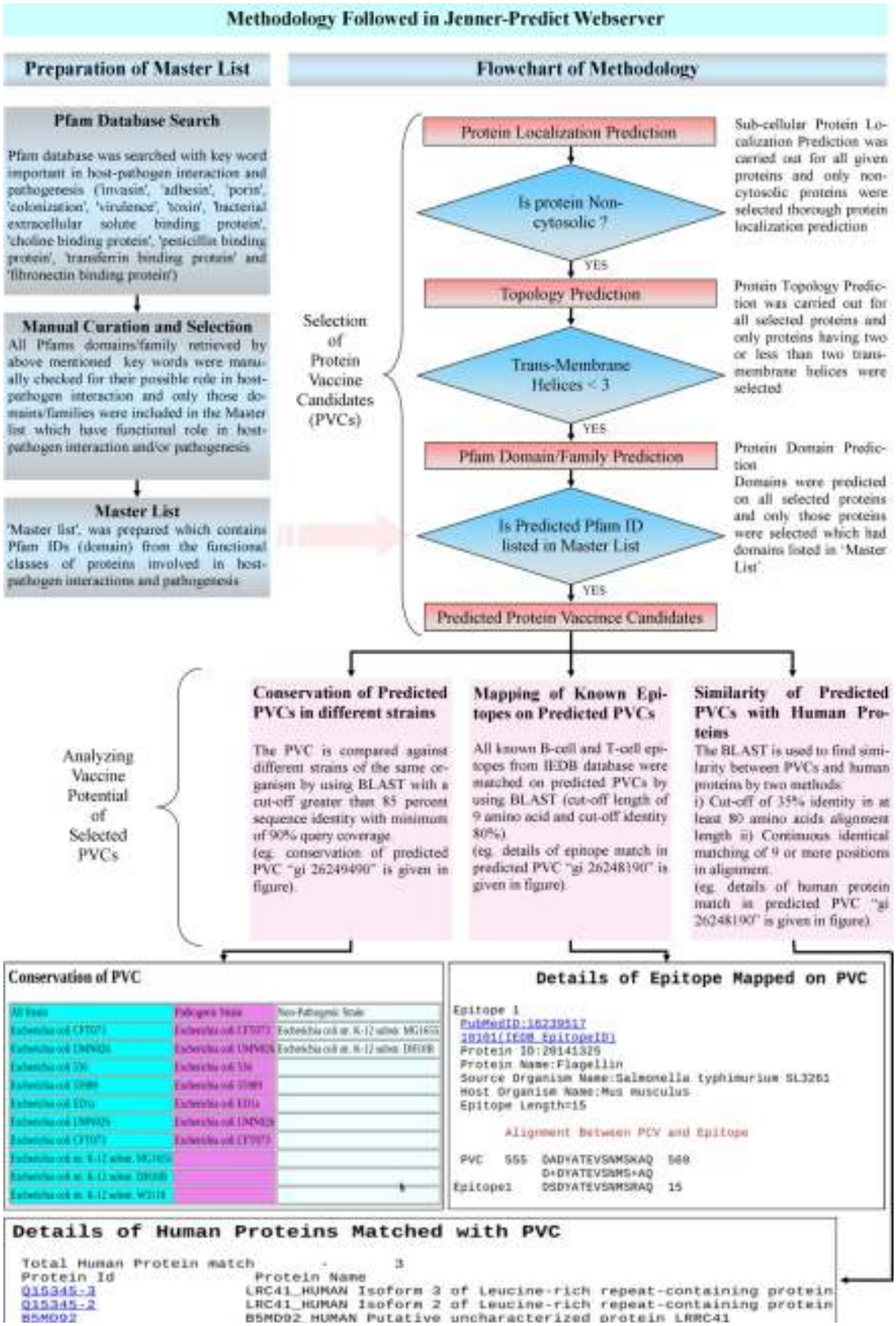


Figure 2.1: Graphical abstract of Jenner-Predict web server

2.1 Introduction

In silico prediction has been proved to be of great importance among various disciplines of life sciences including biomedical research (Tarca *et al*., 2007). The conventional vaccine development methods are time consuming as they require cultivation of pathogenic microorganisms in laboratory conditions and their dissection using microbiological, biochemical and immunological methods in order to identify the components important for immunogenecity. These methods are ineffective in circumstances where the cultivation of bacteria is difficult or impossible. The other limitations arise when the expression of protective antigens is less or absent in *in vitro* conditions compared to *in vivo* diseased conditions (Rappuoli, 2000). With comparison to conventional live attenuated vaccines, subunit vaccines are more reliable as far as safety is concerned (Kimman, 1992). The integration of genomics in vaccine research (vaccinogenomics) is expected to revolutionize novel vaccine candidate identification (Gay *et al*., 2007) which is an essential and important component in subunit vaccine development. Computational approach, especially reverse vaccinology (RV) method assists the identification of vaccine candidates from genomes without culturing microorganisms and thus facilitates the subunit vaccine development. These methods are useful in reducing time, cost and number of wet lab experiments (Rappuoli, 2000).

The RV is a computational pipeline for identification of vaccine candidates against microorganisms from their genome sequences. Thus, all proteins of an organism can be screened computationally for their vaccine potential. Significant success of this principle for vaccine development had already been demonstrated in several pathogens, including *Neisseria meningitidis* (Pizza *et al*., 2000), *Helicobacter pylori* (Chakravarti *et al*., 2000), *Streptococcus pneumoniae* (Wizemann *et al*., 2001), *Porphyromonas gingivalis* (Ross *et al*., 2001), *Chlamydia pneumoniae* (Montigiani *et al*., 2002) and *Bacillus anthracis* (Ariel *et al*., 2002). The relevance of this method was recognized when vaccines developed from capsular polysaccharides of *N. meningitidis* B had failed due to cross reactivity against human tissue (Pizza *et al*., 2000). Application of RV techniques for PVC identification and then *in vivo* testing led to the development of licensed broad specificity protein vaccine, 5CVMB, against *N. meningitidis*. This vaccine contains 5 protein antigen components, GNA2132, GNA1870, GNA1030, GNA2091 and NadA, which were primarily discovered by RV methods (Giuliani *et al*., 2006). However in earlier RV techniques, protein localization (secretory, outer-

membrane, transporter or others) was used as the main criterion for identification of PVCs. As a result, large number of proteins was required to be expressed, purified and tested to obtain few vaccine candidates leading to enormous loss of cost and time.

On the other hand, identification of immunogenic proteins (PVCs) by using epitope prediction software and web servers has several limitations. Comparative studies have shown that B-cell epitopes (BCEs) and class II MHC-binding T-cell epitopes (TCEs) prediction methods are not accurate (Blythe and Flower, 2005; Gowthaman and Agrewala, 2007; Ponomarenko and Bourne, 2007; Zhang *et al*., 2010). Over-prediction, inability in exact position prediction of epitopes and absence of success in identifying known epitopes in proteins are major concerns in vaccine candidate identification. Until now the available PVCs prediction software and web servers have not been much effective for identification of vaccine candidates from genomes for vaccine design. VaxiJen server, based on discriminant analysis and partial least square (DA-PLS) methods, was developed by using datasets of known (positive) protective antigenic and non-antigenic (negative) proteins to predict PVCs (Doytchinova and Flower, 2007). Surprisingly, it predicted more than half of proteins from a given bacterial proteome as protective antigens with default parameters making its usage almost impractical. Further, existing software and web servers predict different proteins as vaccine candidates from same proteome sequences. For example, different proteins were predicted from *S. pneumoniae* proteome by VaxiJen server (Doytchinova and Flower, 2007) and new enhanced reverse vaccinology environment (NERVE) (Vivona *et al*., 2006) software. From 2202 proteins of *S. pneumoniae*, VaxiJen (with cut-off of 0.6) and NERVE predicted 313 and 58 as PVCs, respectively while only 20 proteins were common between them. None of the common PVCs matched with 18 known vaccine candidates in *S. pneumoniae* (Table 2.1). This outcome complicates decision process regarding which tool's output should be taken for experimental testing to identify vaccine candidates. The method used in NERVE (Vivona *et al*., 2006) and Vaxign (He *et al*., 2010) tools presumed that extracellular proteins having adhesin-likeliness are potential vaccine candidates. Although adhesin-likeliness of a protein is an important criterion, it should not be considered as the only one because several non-adhesin functional classes of proteins (*i.e.* invasin, porin, flagellin, etc.) are also involved in host-pathogen interactions or pathogenesis and many of them are known to be antigenic (Cao *et al*., 2011; Chen *et al*., 2006; Easton *et al*., 2005; Ko and Splitter, 2003; Potter *et al*., 1999; Rappuoli *et al*., 1996; Schorey *et al*., 1996; Tang and Holden, 1999; Tong *et al*., 2005; Turbyfill *et al*., 2008; Wizemann *et al*., 1999; Zarantonelli

et al., 2006; Zou *et al*., 2010). It has been suggested that targeting host-pathogen interactions and disease processes at molecular level can be used for novel vaccine discovery (Gay *et al*., 2007). In several cases, the immune responses against these non-adhesins were known to provide protection against microbial infection (Cao *et al*., 2011; Chen *et al*., 2006; Easton *et al*., 2005; Ko and Splitter, 2003; Potter *et al*., 1999; Rappuoli *et al*., 1996; Schorey *et al*., 1996; Tang and Holden, 1999; Tong *et al*., 2005; Turbyfill *et al*., 2008; Wizemann *et al*., 1999; Zarantonelli *et al*., 2006; Zou *et al*., 2010). Invasin, porin, flagellin and toxin have roles in host cell invasion (Palumbo and Wang, 2006); transportation activity associated with pathogenesis and virulence (Achouak and Heulin, 2001); chemotaxis, adhesion and colonization making pathogenic bacteria to be virulent (Ramos *et al*., 2004); and host cell death (Galán, 2005), respectively. Bacterial fibronectin-binding proteins (FBPs) target host fibronectin for adhesion and colonization (Henderson *et al*., 2011); transferrin-binding proteins (TBP) are used by bacteria to obtain iron directly from host transferrins (Ratledge and Dover, 2000); and penicillin-binding proteins (PBPs) are involved in peptidoglycan biosynthesis to maintain cell wall structure and protection (Sauvage *et al*., 2008). The solute binding proteins (SBPs) are used to capture nutrients like iron to overcome the environment devoid of free nutrients within the host (Zou *et al*., 2010). Choline-binding proteins (CBPs) in some bacteria perform adhesin-like function (Rosenow *et al*., 1997). Functional classes of proteins involved in virulence (Tang and Holden, 1999), invasion (Turbyfill *et al*., 2008) and colonization (Tong *et al*., 2005); porins (Easton *et al*., 2005) and flagellin (Chen *et al*., 2006); and binding proteins of choline (Cao *et al*., 2011), penicillin (Zarantonelli *et al*., 2006), transferrin (Potter *et al*., 1999), fibronectin (Schorey *et al*., 1996) and solute (Zou *et al*., 2010) are important in host-pathogen interactions and pathogenesis. Since many proteins from these functional classes provide protective immune responses against microbial infection (Cao *et al*., 2011; Chen *et al*., 2006; Easton *et al*., 2005; Ko and Splitter, 2003; Potter *et al*., 1999; Rappuoli *et al*., 1996; Schorey *et al*., 1996; Tang and Holden, 1999; Tong *et al*., 2005; Turbyfill *et al*., 2008; Wizemann *et al*., 1999; Zarantonelli *et al*., 2006; Zou *et al*., 2010), the knowledge of host-pathogen interactions related to bacterial pathogenesis could be used to rationalize and improve vaccine candidate prediction.

A web server, Jenner-Predict, has been developed which is capable of predicting PVCs from proteome/protein sequences. It is based on the principle that non-cytosolic proteins having functions (domains) important in host-pathogen interactions and/or pathogenesis are potential vaccine candidates (Figure 2.1). It has two broad components: PVCs prediction and

analysis of their vaccine potential. The PVCs prediction is performed in three sequential steps: prediction of subcellular localization, expressibility in laboratory and presence of domains critical in host-pathogen interactions and pathogenesis. Software PSORTb 3.0 is used for protein subcellular localization prediction (Nancy *et al*., 2010). A protein has high probability of failure to express in experiment (Pizza *et al*., 2000) when it has more trans-membrane helices. HMMTOP 2.0 (Tusnady and Simon, 2001) software is used for topology prediction and proteins with more than two trans-membrane helices are discarded. Proteins pass through above two filters and having domains involved in host-pathogen interactions and pathogenesis from functional classes of adhesin, invasin, toxin, porins, colonization, virulence, flagellin, penicillin-binding, choline-binding transferring-binding, fibronectin-binding and solute-binding proteins are selected as vaccine candidates. Standalone Pfam sequence search is used for prediction of domains (Punta *et al*., 2012). Vaccine potential of PVCs is predicted on the basis of their possible immunogenicity, absence of autoimmunity, and conservation across different pathogenic and non-pathogenic strains of same bacteria. Known BCEs and TCEs from immune epitope database (IEDB) (Vita *et al*., 2010) are mapped separately on predicted PVCs to know their possible immunogenic region and immunogenic potential. This mapping of antigenic determinant (epitope) is instrumental in predicting humoral (BCE) or cellular (TCE) or both immune responses of PVC. Since PVCs specific to pathogenic strains are expected to be involved in virulence (Tang and Holden, 1999), therefore conservation of PVCs in different pathogenic strains of same organism is determined to provide more robust vaccine candidates. The PVCs having homolog(s) in host (human) are provided by the web server. Such PVCs may produce autoimmunity (Iwai *et al*., 2005) or less immune response (Grossman and Paul, 2001). Taking into account above criteria, output of the web server is provided as prioritized PVCs in the result table. Comparison among PVC prediction methods has shown that Jenner-Predict server's performance is better.

2.2 Methods

2.2.1 *Data collection and generation*

Proteomes of all pathogenic and non-pathogenic bacteria were taken from NCBI ftp (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.faa.tar.gz). The proteomes of *S. pneumoniae* strain 70585 and *Escherichia coli* uropathogenic strain CFT073 were collected from above

proteomes. Human proteome sequences were also downloaded from the EBI ftp site (ftp://ftp.ebi.ac.uk/pub/databases/integr8/fasta/proteomes) for prediction of human homologs in predicted PVCs. For the development of web server, standalone version of four softwares (Figure 2.1), NCBI BLAST (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/), PSORTb 3 (http://www.psort.org/psortb/), HMMTOP 2.0 (http://www.enzim.hu/hmmtop/) and HMMER 3.0 (http://hmmer.janelia.org/) were downloaded from their respective websites. PSORTb 3 predicts subcellular localization of a given protein sequence based on its amino acid composition, similarity to proteins of known localization, and presence of different motifs and signal peptides (Nancy *et al*., 2010). HMMTOP software uses hidden Markov model to predict transmembrane helices based on the difference in the amino acid distributions in various structural parts of proteins (Tusnady and Simon, 2001). For prediction of domains in protein sequences, Perl program, pfam_scan.pl and Pfam library of hidden Markov models (HMMs) for protein families were downloaded from Pfam website (http://pfam.janelia.org/).

For prediction of immunogenic regions in PVCs, experimentally known immunogenic epitope sequences of all TCE and BCE assays were downloaded from IEDB in CSV format. Peptide epitope with literature reference, epitope ID, GI of source protein, and source and host organism's information were extracted from these TCEs and BCEs assays. In case of TCEs, MHCs allele names were also extracted. For discontinuous BCEs, corresponding protein sequences were downloaded from database, and stretch of continuous sub-part protein containing all the residues of discontinuous epitope positions was extracted. These subsequences were stored in 'fasta' format for comparison against predicted PVCs.

2.2.2 ***Collection of data for web server validation***

Experimentally known protective antigens were collected from four diverse sources to evaluate the performance of Jenner-Predict server against existing methods. Known non-cytosolic protective PVCs from the two pathogenic bacteria, *S. pneumoniae* (gram-positive) and *E. coli* (gram-negative) were collected from literature (Table 2.1 and 2.2). Different experiments had identified 18 and 28 non-cytosolic proteins to be protective antigens in *S. pneumoniae* and *E. coli*, respectively (Table 2.1 and 2.2). To demonstrate the effectiveness of web server in predicting vaccine candidates across bacteria, non-cytosolic protective antigens sequences reported in 'Protegen' database (Yang *et al*., 2011) were retrieved for evaluation as well. Out of the 257 reported bacterial protective PVCs, 211 were predicted to be non-

cytosolic by PSORTb. After removing 11 antigens having more than 2 trans-membrane helices and sequences which are 90 percent identical among themselves by using CD-HIT (http://weizhong-lab.ucsd.edu/cd-hit/ref.php), 177 bacterial protective PVCs from Protegen database were selected for evaluation. In addition to above, non-cytosolic proteins from datasets used for VaxiJen (Doytchinova and Flower, 2007) server development were also taken for evaluation. Positive and negative training and test datasets containing 100 sequences of each in the form of Swiss-Prot IDs were collected and then their sequences were retrieved. PSORTb was used to predict their localization and only non-cytosolic proteins were retained. Finally, 83 and 33 non-cytosolic positive (protective antigen) and negative (non-antigen) sequences were selected for comparison of performances. The sequences used for validation in both Protegen and VaxiJen datasets are highly diverse and more than 90% sequences are less than 40% identical.

Table 2.1: Protein vaccine candidates (PVCs) reported in *S.pneumoniae*

S.No.	Name of gene/protein	Gene ID	Localization	Ref.
1.	Pneumolysin (Thiol-activated cytolysin)	225859688	Extracellular	(Alexander *et al*., 1994)
2.	Pneumococcal choline binding protein A (PcpA)	225859909	Unknown	(Glover *et al*., 2008)
3.	BVH-3 PhpA protein	225858797	Unknown	(Hamel *et al*., 2004)
4.	Autolysin lytA	225859701	Extracellular	(Berry *et al*., 1989)
5.	Endo-beta-N-acetylglucosaminidase (SP046)	225858758	Extracellular	(Wizemann *et al*., 2001)
6.	1,4-beta-N-acetylmuramidase (SP091)	225859330	Extracellular	(Wizemann *et al*., 2001)
7.	PspA	225857997	Extracellular	(Yamamoto *et al*., 1997)
8.	ABC transporter permease (Pit)	225858856	Cytoplasmic Membrane	(Brown *et al*., 2001)
9.	Histidine triad protein B (SP036)	225858962	Non-Cytoplasmic	(Wizemann *et al*., 2001)
10.	Putative protease maturation protein A	225858774	Cytoplasmic Membrane	(Overweg *et al*., 2000)

	(PpmA)			
11.	PsaA	225859406	Cytoplasmic Membrane	(Briles *et al*., 2000)
12.	Pneumococcal vaccine antigen A (SP101)	225858817	Cytoplasmic Membrane	(Wizemann *et al*., 2001)
13.	Serine/threonine protein kinase (StkP)	225859485	Cytoplasmic Membrane	(Giefing *et al*., 2008)
14.	Pneumoniae neuraminidase (NanA)	225859446	Cell wall	(Tong *et al*., 2005)
15.	CbpA or PspC or Hic or SpsA	225858707	Cytoplasmic Membrane	(Ogunniyi *et al*., 2001)
16.	Zinc metalloprotease (ZmpB)	225858492	Cell wall	(Gong *et al*., 2011)
17.	Endo-alpha-N-acetylgalactosaminidase	225858223	Cell wall	(Caines *et al*., 2008)
18.	Pullulanase	225858118	Cell wall	(Bongaerts *et al*., 2000)

Table 2.2: Protein vaccine candidates (PVCs) reported in *E. coli*

S.No.	Locus	Gene Name	Gene ID (GI)	Cellular Localization	Ref.
1.	c0185	FhuA	26246096	OuterMembrane	(Hagan and Mobley, 2007)
2.	c0214	YaeT	26246123	OuterMembrane	(Hagan and Mobley, 2007)
3.	c0652	OmpT	26246544	OuterMembrane	(Hagan and Mobley, 2007)
4.	c0900	OmpX	26246790	OuterMembrane	(Hagan and Mobley, 2007)
5.	c1071	OmpF	26246956	OuterMembrane	(Hagan and Mobley, 2007)
6.	c1093	OmpA	26246978	OuterMembrane	(Hagan and Mobley, 2007)
7.	c1250	IroN	26247124	OuterMembrane	(Hagan and Mobley, 2007)
8.	c3655	Ag43	26249490	OuterMembrane	(Hagan and Mobley, 2007)
9.	c1560	NmpC	26247429	OuterMembrane	(Hagan and Mobley, 2007)
10.	c1722	OmpW	26247587	OuterMembrane	(Hagan and Mobley, 2007)
11.	c2187	YeaF	26248041	OuterMembrane	(Hagan and Mobley, 2007)
12.	c2338	FliC	26248190	Extracellular	(Hagan and Mobley, 2007)
13.	c2482	colicin	26248334	OuterMembrane	(Hagan and Mobley, 2007)
14.	c2758	OmpC	26248604	OuterMembrane	(Hagan and Mobley, 2007)
15.	c3610	Iha	26249445	OuterMembrane	(Hagan and Mobley, 2007)
16.	c3623	IutA	26249458	OuterMembrane	(Hagan and Mobley, 2007)
17.	c3781	TolC	16148612	OuterMembrane	(Hagan and Mobley, 2007)

			1		
18.	c4095	YheE	26249919	Cytoplasmic Membrane	(Hagan and Mobley, 2007)
19.	c4308	ChuA	26250130	OuterMembrane	(Hagan and Mobley, 2007)
20.	c4894	Tsx	26250708	OuterMembrane	(Hagan and Mobley, 2007)
21.	c4929	BtuB	229106362	OuterMembrane	(Hagan and Mobley, 2007)
22.	c5006	LamB	26250818	OuterMembrane	(Hagan and Mobley, 2007)
23.	c5174	IreA	26250982	OuterMembrane	(Hagan and Mobley, 2007)
24.	c5400	FimH	26251208	Unknown	(Denich *et al*., 1991)
25.	c5188	PapA	26250996	Extracellular	(Langermann *et al*., 2000)
26.	c3389	_	26249224	Cytoplasmic Membrane	(Durant *et al*., 2007)
27.	c0393	_	26246291	Unknown	(Durant *et al*., 2007)
28.	c4424	_	26250246	OuterMembrane	(Durant *et al*., 2007)

2.2.3 ***Server architecture***

The web server comprised of a client interface and a main application program. The client interface was developed using HTML language which takes input either in the form of protein sequence(s) in fasta format or a proteome of listed bacteria. The submitted fasta sequence(s)/proteome are processed by the in-house backend Perl-CGI script which posts information provided by the user to the main application program in a queue format. This Perl-CGI script generates an URL link where the status information or output of a given job will be available. The self developed programs and other available standalone software (Figure 2.1) are used by 'main application' program for the analysis of protein sequences one after another to predict PVCs. The main application program also provides the output table in the form of prioritized PVCs.

2.2.4 ***Pfam domain selection***

Domains are basic building blocks of proteins. Searching of a protein sequence against Pfam library of HMMs enables to find domain architecture present in that protein (Punta *et al*., 2012). The Pfam has been used in several genome projects including human for large scale functional annotation of genomic data (Lander *et al*., 2001). A list called, 'Master list', was prepared which contains Pfam IDs (domain) from the functional classes of proteins

involved in host-pathogen interactions and pathogenesis (Cao *et al*., 2011; Chen *et al*., 2006; Easton *et al*., 2005; Ko and Splitter, 2003; Potter *et al*., 1999; Rappuoli *et al*., 1996; Schorey *et al*., 1996; Tang and Holden, 1999; Tong *et al*., 2005; Turbyfill *et al*., 2008; Wizemann *et al*., 1999; Zarantonelli *et al*., 2006; Zou *et al*., 2010). For preparing the list, Pfam database was subjected to text search with individual key words 'invasin', 'adhesin', 'porin', 'colonization', 'virulence', 'toxin', 'bacterial extracellular solute binding protein', 'choline binding protein', 'penicillin binding protein', 'transferrin binding protein' and 'fibronectin binding protein' to identify domains from each classes of proteins. Then all hits of domains from each keyword were manually checked for their possible role in host-pathogen interactions. Only those families/domains were included in the 'Master list' which have significant functional role in host-pathogen interactions and/or pathogenesis (Table 2.3). This 'Master list' of domains was used for the prediction of PVCs from non-cytosolic proteins (Figure 2.1).

Table 2.3: Key words used and selection of Pfam domains for protein vaccine candidate prediction

S. No.	Key word used for PFam domain search	No. of domain hits	No. of selected domains*	Reference
1.	Adhesin	166	96	(Wizemann *et al*., 1999)
2.	Choline binding protein	29	12	(Cao *et al*., 2011)
3.	Bacterial extracellular solute-binding protein	36	8	(Zou *et al*., 2010)
4.	Porin	66	46	(Easton *et al*., 2005)
5.	Invasin	30	25	(Turbyfill *et al*., 2008)
6.	Fibronectin-binding protein	50	25	(Schorey *et al*., 1996)
7.	Transferrin-binding protein	24	6	(Potter *et al*., 1999)
8.	Virulence	402	145	(Tang and Holden, 1999)
9.	Penicillin-binding Protein	14	8	(Zarantonelli *et al*., 2006)
10.	Flagellin	22	12	(Chen *et al*., 2006)
11.	Colonization	23	14	(Tong *et al*., 2005)
12.	Host-pathogen interaction	9	4	(Ko and Splitter, 2003)

13.	Toxin	542	110	(Rappuoli *et al*., 1996)

*Only those families/domains were included which are involved in host-pathogen interactions and/or pathogenesis

2.2.5 ***Implementation***

The server, Jenner-Predict, has two major components: PVCs prediction and analysis of their vaccine potential (Figure 2.1). Software PSORTb 3.0 (Nancy *et al*., 2010) and HMMTOP 2.0 (Tusnady and Simon, 2001) are used to predict subcellular localization and number of transmembrane helices, respectively. The former discards cytoplasmic proteins whereas the latter rejects proteins having more than two transmembrane helices (Pizza *et al*., 2000). Proteins passing through the above two filters are then subjected to Pfam domain/family search to determine their domains. Finally, proteins matching with Pfam domains/families listed in the 'Master list' (Table 2.3) are selected as PVCs.

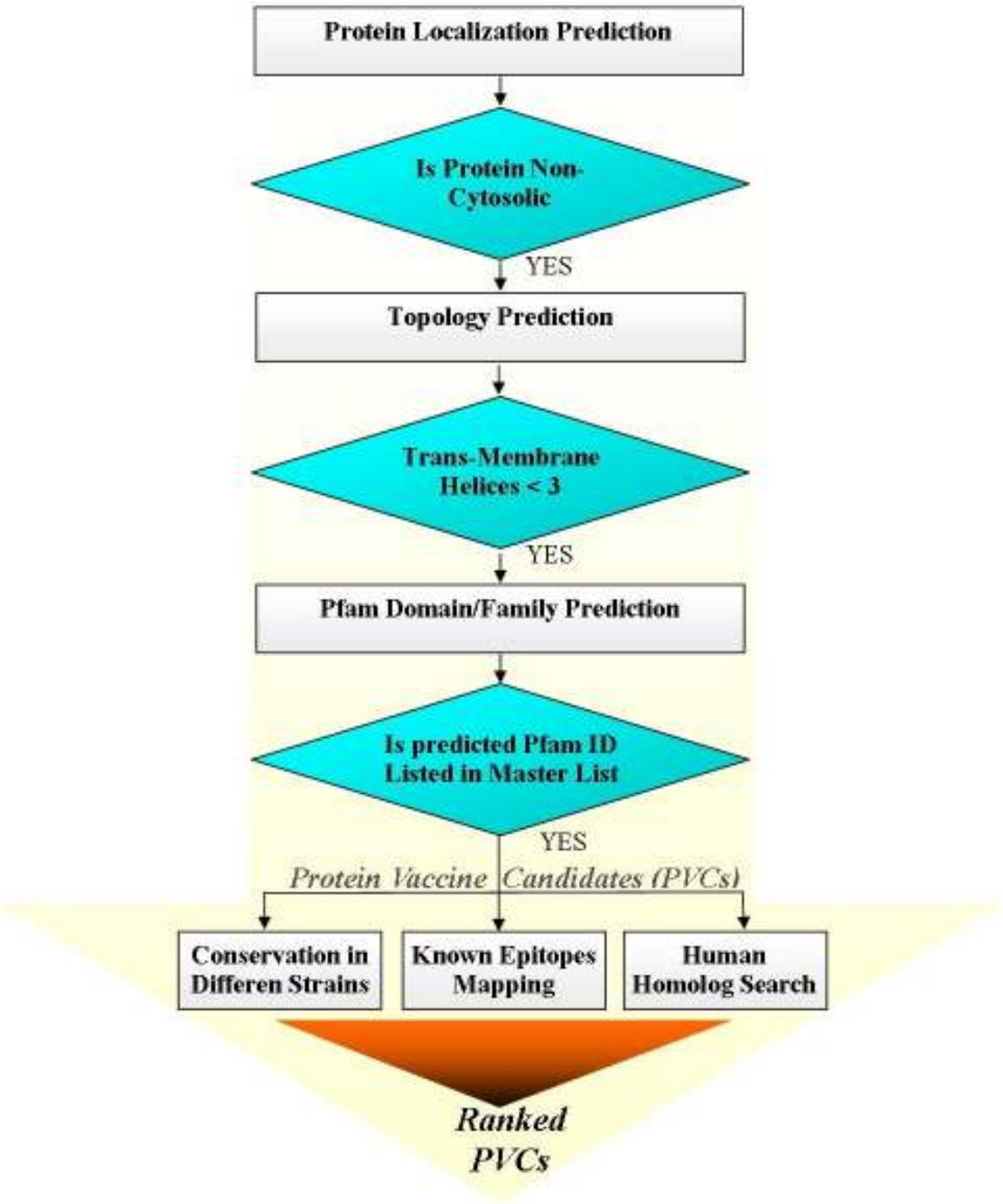


Figure 2.2: Flow chart depicting methodology of Jenner-Predict web server

Vaccine potential of the predicted PVCs' is performed by taking three different measures into account: immunogenicity, autoimmunity and conservation (Figure 2.1). Immunogenic potential (putative immunogenic regions) of PVCs is predicted by matching of IEDB epitopes against the PVCs by using standalone BLAST with minimum matching length of 9 (Del Val *et al*., 1991) and 80% identity cut-off. For autoimmunity prediction, the BLAST is used to

find similarity between PVCs and human proteins by two different methods: i) cut-off of 35% identity in at least 80 amino acids length of PVC (Fiers *et al*., 2004), and ii) continuous identical matching of 9 or more positions in the alignment (Del Val *et al*., 1991). BLAST is also used to identify conservation of PVCs in different pathogenic strains of a given organism. The PVC is compared against different strains of the same organism with a cut-off greater than 85 percent sequence identity with minimum of 90% query coverage. To determine conservation of PVC in pathogenic and non-pathogenic strains separately, names of pathogenic and non-pathogenic strains of each organism are stored in two separate flat files under each category. Information on pathogenic or non-pathogenic strains of each individual organism is extracted from the respective files.

2.3 Results

The web server, Jenner-Predict, has been developed to predict PVCs from proteome or protein(s) sequences for subunit vaccine development on the basis of domains critical to host-pathogen interactions and pathogenesis. Besides predicting PVCs, it also furnishes information crucial to determine their vaccine capability in terms of immunogenic potential by matching PVCs against IEDB epitopes, autoimmunity through matching PVCs with human proteome, and their conservation across different pathogenic and non-pathogenic strains of the organism. A tutorial explaining how to submit a job as well as user-friendly interpretation of results is available at the web server's home page. The web server gives higher priority to PVCs containing more IEDB epitope matches as they increase their possibility to be immunogenic. The PVCs containing identical IEDB epitopes match are shown in white background. The web server also decreases priority of PVCs having human homologs as such PVCs should be discouraged from further vaccine development process. This prioritization is instrumental in selecting few PVCs for further vaccine development experiments. The performance of the web server was evaluated against reported vaccine candidates in *S. pneumoniae* (gram positive) and *E. coli* (gram negative), proteins (both positive and negative) used for development of VaxiJen server (Doytchinova and Flower, 2007) and protective antigens from more than 40 bacteria reported in Protegen database (Yang *et al*., 2011).

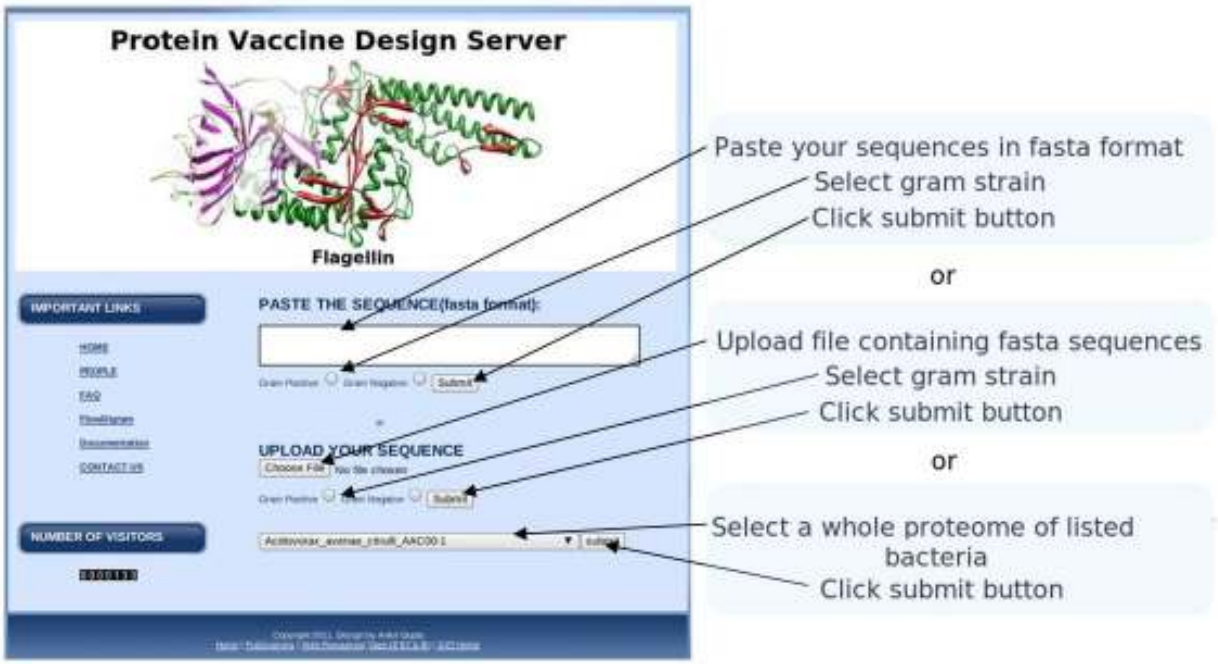


Figure 2.3: Job submission web page of Jenner-Predict depicting three different methods for submission of a job

2.3.1 ***PVCs prediction in S. pneumoniae and E. coli***

In *S. pneumoniae* proteome, Jenner-Predict server predicted 69 proteins as vaccine candidates (Figure2.4 and http://14.139.240.55/vaccine/results_data/98%5C98_58892.txt.pri.html). As VaxiJen server predicts more than half of a proteome as vaccine candidates in any bacteria with default VaxiJen probability score, 0.4, a cut-off of 0.6 was considered to restrict the number of PVCs so that the performance of this approach can be compared against all other methods. From *S. pneumoniae* proteome, our web server predicted 10 out of 18 known non-cytoplasmic PVCs whereas the software, NERVE, and servers, Vaxign and VaxiJen, predicted only 7, 6 and 3 PVCs, respectively (Figure 2.4 and Table 2.4). As compared to other methods, the PVCs predicted exclusively by Jenner-Predict server were STK (Giefing *et al*., 2008), NanA (Turbyfill *et al*., 2008), and PsaA (Talkington *et al*., 1996) which are having PASTA, BNR, and SBP domains, respectively. Other than 10 reported vaccine candidates from 69 PVCs (Figure 2.4), our web server predicted 18 ABC-transporters and solute-binding, 6 choline-binding, 4 penicillin-binding, 2 LysM-containing domain, 3 cell wall anchors, etc. ABC-transporter (Garmory and Titball, 2004), choline-binding (Cao *et al*., 2011), and penicillin-binding (Zarantonelli *et al*., 2006) proteins are known to be potential PVCs in different bacteria including *S. pneumoniae*.

In *E. coli* proteome, Jenner-Predict server predicted 253 proteins as vaccine candidates (Figure 2.5) whereas the NERVE and Vaxign predicted more than 500 and 280 proteins, respectively (Figure 2.5). Our web server predicted 23 out of 28 known PVCs whereas software, NERVE and servers, Vaxign and VaxiJen, predicted 21, 18 and 21 PVCs, respectively. The PVCs missed out by other methods due to being non-adhesins were OmpA (outer-membrane protein A), BtuB (cobalamin outer-membrane transporter), TolC (channel protein) and IreA (putative iron-regulated outer membrane virulence proteins) (Table 2.5). Besides 23 known protective antigens, the majority of predicted PVCs by our web server were 51 BPD transporter proteins, 32 solute-binding proteins (Zou *et al*., 2010) and 32 fimbrial proteins (Sadilkova *et al*., 2012).

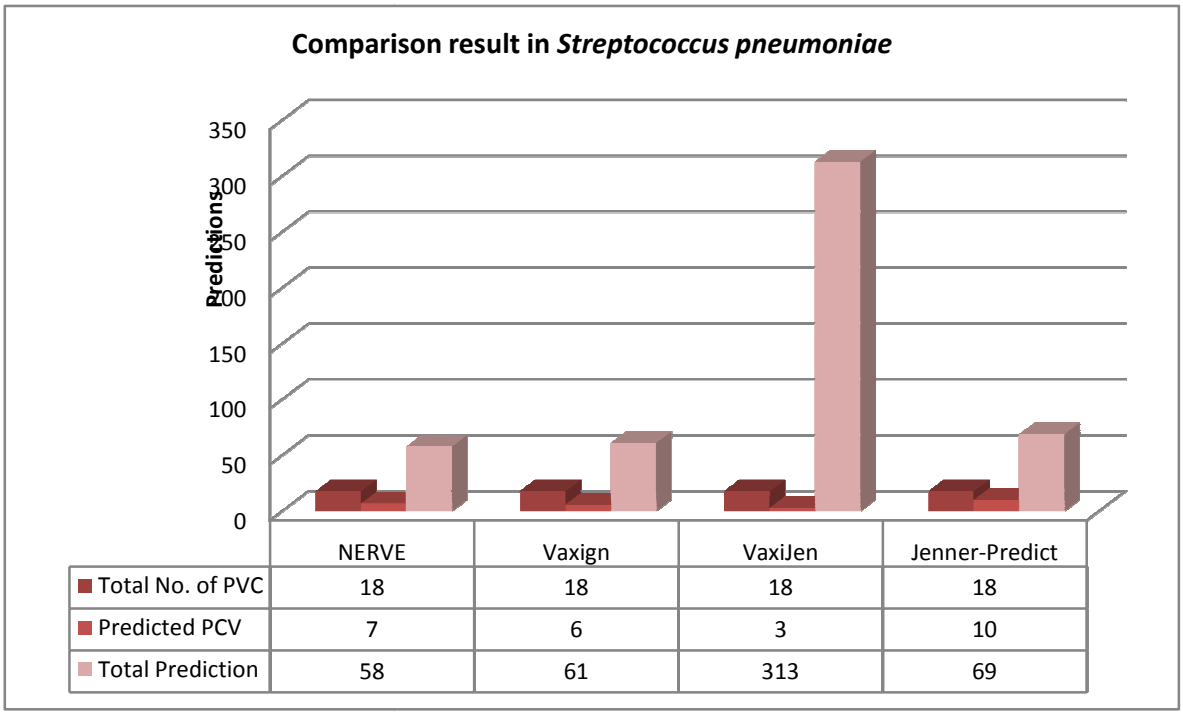


Figure 2.4: Comparative results of predicted PVCs in *Streptococcus pneumoniae* through different methods.

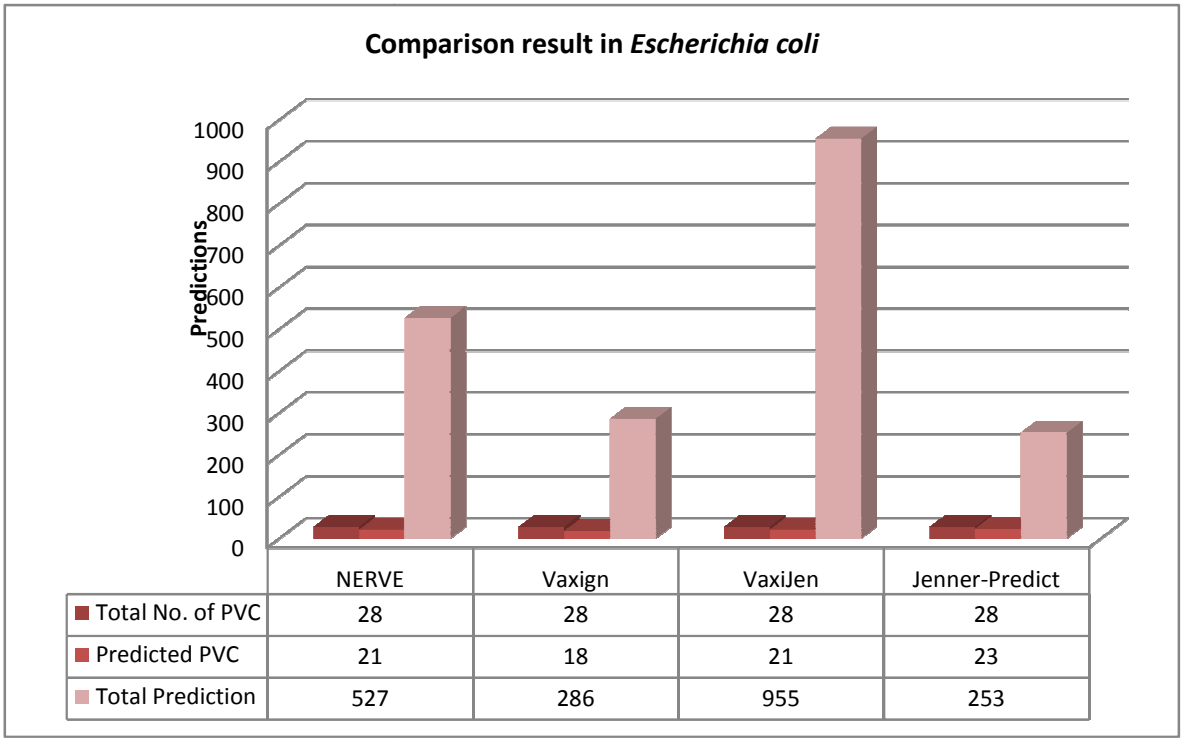


Figure 2.5: Comparative results of predicted PVCs in *Escherichia coli* through different methods.

Table 2.4: Detailed comparison of results for predicted protein vaccine candidate (PVC) by software, NERVE, and web servers, Vaxign, VaxiJen and Jenner-Predict from *Streptococcus pneumoniae* 70585 (gram-positive) against experimentally known protective antigens*

#S. No.	Name of gene/protein	Gene ID	Localization	Nerve	Vaxign	VaxiJen	Jenner-Predict
1.	Pneumolysin (Thiol-activated cytolysin)	225859688	Extracellular	YES	NO	NO	YES
2.	Pneumococcal choline binding protein A (PcpA)	225859909	Unknown	YES	YES	NO	YES
3.	BVH-3 PhpA protein	225858797	Unknown	YES	NO	NO	NO
4.	Autolysin lytA	225859701	Extracellular	YES	YES	NO	YES
5.	Endo-beta-N-acetylglucosaminidase (SP046)	225858758	Extracellular	YES	YES	NO	YES

6.	1,4-beta-N-acetylmuramidase	225859330	Extracellular	YES	YES	NO	YES
7.	PspA	225857997	Extracellular	NO	YES	NO	YES
8.	ABC transporter permease (Pit)	225858856	Cytoplasmic Membrane	NO	NO	NO	NO
9.	Histidine triad protein B (SP036)	225858962	Non-Cytoplasmic	NO	NO	NO	NO
10.	Putative protease maturation protein A (PpmA)	225858774	Cytoplasmic Membrane	NO	NO	YES	NO
11.	PsaA	225859406	Cytoplasmic Membrane	NO	NO	NO	YES
12.	Pneumococcal vaccine antigen A (SP101)	225858817	Cytoplasmic Membrane	NO	NO	NO	NO
13.	Serine/threonine protein kinase (StkP)	225859485	Cytoplasmic Membrane	NO	NO	NO	YES
14.	Pneumoniae neuraminidase (NanA)	225859446	Cell wall	NO	NO	YES	YES
15.	CbpA or PspC or Hic or SpsA	225858707	Cytoplasmic Membrane	NO	NO	YES	NO
16.	Zinc metalloprotease (ZmpB)	225858492	Cell wall	NO	NO	NO	NO
17.	Endo-alpha-N-acetylgalactosaminidase	225858223	Cell wall	YES	YES	NO	YES
18.	Pullulanase	225858118	Cell wall	NO	NO	NO	NO

* See details in methods section. Jenner-Predict server is based on domains involved in host-pathogen interactions which are important in pathogenesis and disease establishment. For comparison with VaxiJen, a cut-off of 0.6 was used instead of default parameter 0.4 as it predicts almost half of proteome as vaccine candidates with default parameter.

S. No. indicates Serial Number; YES or NO denotes the corresponding protein is predicted or not, respectively by the corresponding software or web server.

Table 2.5: Detailed comparison of results for predicted protein vaccine candidate (PVC) by software, NERVE, and web servers, Vaxign, VaxiJen and Jenner-Predict from *Escherichia coli* Uropathogenic strain CFT073 (gram-negative) against experimentally known protective antigens*

#S. No.	Locus	Gene Name	Gene ID (GI)	Cellular Localization	NERVE	Vaxign	VaxiJen	Jenner-Predict
1.	c0185	FhuA	26246096	OuterMembrane	YES	YES	NO	YES
2.	c0214	YaeT	26246123	OuterMembrane	YES	NO	NO	YES
3.	c0652	OmpT	26246544	OuterMembrane	NO	NO	NO	NO
4.	c0900	OmpX	26246790	OuterMembrane	YES	YES	YES	YES
5.	c1071	OmpF	26246956	OuterMembrane	YES	NO	YES	YES
6.	c1093	OmpA	26246978	OuterMembrane	NO	NO	YES	YES
7.	c1250	IroN	26247124	OuterMembrane	YES	YES	YES	YES
8.	c3655	Ag43	26249490	OuterMembrane	YES	YES	YES	YES
9.	c1560	NmpC	26247429	OuterMembrane	YES	YES	YES	YES
10.	c1722	OmpW	26247587	OuterMembrane	YES	YES	YES	YES
11.	c2187	YeaF	26248041	OuterMembrane	YES	YES	NO	NO
12.	c2338	FliC	26248190	Extracellular	YES	YES	YES	YES
13.	c2482	colicin	26248334	OuterMembrane	YES	YES	YES	YES
14.	c2758	OmpC	26248604	OuterMembrane	YES	YES	YES	YES
15.	c3610	Iha	26249445	OuterMembrane	YES	YES	YES	YES
16.	c3623	IutA	26249458	OuterMembrane	YES	YES	YES	YES
17.	c3781	TolC	161486121	OuterMembrane	NO	NO	NO	YES
18.	c4095	YheE	26249919	Cytoplasmic Membrane	NO	NO	YES	NO
19.	c4308	ChuA	26250130	OuterMembrane	YES	YES	YES	YES
20.	c4894	Tsx	26250708	OuterMembrane	YES	YES	YES	NO

21.	c4929	BtuB	229106362	OuterMembrane	NO	NO	NO	YES
22.	c5006	LamB	26250818	OuterMembrane	YES	YES	YES	YES
23.	c5174	IreA	26250982	OuterMembrane	NO	NO	YES	YES
24.	c5400	FimH	26251208	Unknown	YES	YES	YES	YES
25.	c5188	PapA	26250996	Extracellular	YES	YES	YES	YES
26.	c3389	__	26249224	Cytoplasmic Membrane	NO	NO	NO	NO
27.	c0393	__	26246291	Unknown	YES	YES	YES	YES
28.	c4424	__	26250246	OuterMembrane	YES	YES	YES	YES

* See details in methods section. Jenner-Predict server has been developed by us and is based on domains involved in host-pathogen interactions which are important in pathogenesis and disease establishment. For comparison with VaxiJen, a cut-off of 0.6 was used instead of default parameter 0.4 as it predicts almost half of proteome as vaccine candidates with default parameter.

S. No. indicates Serial Number; YES or NO denotes the corresponding protein is predicted or not, respectively by the corresponding software or web server.

2.3.2 ***Prediction of PVCs against protegen database and datasets used in VaxiJen server development***

The results of our web server for prediction of known protective vaccine candidates from more than 40 diverse bacteria reported in Protegen database and its comparison against other similar methods has been presented in Figure 2.6. Our web server predicted 137 out of 177 protective antigens (PAs) from Protegen database whereas software, NERVE, and servers, Vaxign and VaxiJen, predicted 121, 89 and 97, respectively. The PAs which were only predicted by our method and skipped by others (NERVE, Vaxign and VaxiJen) belong to functional classes of solute binding, toxin, invasin, etc (Table 2.6). The Jenner-Predict server was found to be efficient in discriminating between antigens and non-antigens. From the 83 Pas (positive dataset) used in VaxiJen server development, NERVE, Vaxign and VaxiJen predicted 53, 47 and 46 proteins, respectively whereas our web server predicted 59 PVCs (Figure 2.7 and Table 2.7). From negative dataset (considered as non-antigens) of 33 proteins, the NERVE, Vaxign and VaxiJen methods predicted 8, 5 and 3 proteins to be vaccine candidates, respectively compared to 2 proteins (Q48919 and Q53247) by our web

server (Figure 2.7 and Table 2.8). Negative dataset proteins were considered as non-antigens. But one out of two PVCs predicted by our web server has already been known to be antigenic: fibronectin-attachment protein (Q48919) provides protective immunity against *Mycobacterium avium* infection (Lee *et al*., 2009).

2.3.3 ***Validation of Jenner-Predict***

Sensitivity and specificity indices of different PVC prediction methods have been presented in Figure 2.7. Jenner-Predict server's PVCs prediction accuracy is better at all levels: bacterial proteomes, Protegen database and datasets used for VaxiJen server development. Unavailability of total number of known vaccine candidates in a proteome prevented us to calculate sensitivity and specificity values for proteome sequences. The results of PVC prediction from two proteomes by different methods have been provided in Figure 2.4 and Figure 2.5. Detailed comparison of results w.r.t known vaccine candidates of *S. pneumoniae* and *E. coli* by different methods has been provided in Table 2.4 and Table 2.5, respectively. On dataset used in development of VaxiJen server, sensitivity and specificity of our tool were 0.711 and 0.940, respectively whereas comparable methods NERVE, Vaxign and VaxiJen have corresponding values 0.639 and 0.765; 0.494 and 0.853 and 0.554 and 0.909, respectively. For the Protegen database, only sensitivity was calculated as specificity calculation was not feasible due to lack of negative dataset. The sensitivities of NERVE, Vaxign and VaxiJen were 0.684, 0.491 and 0.548, respectively as compared to 0.774 for Jenner-Predict server (Table 2.6A). Further to access the significance of result, the performance of different comparative methods was also carried out on randomly generated datasets of 25%, 50%, and 75% of validated Protegen database PVCs with five replicates for each set. Performance of each comparative method is provided in Table 2.6B. Performance of Jenner-Predict method was the best and consistent as the sensitivity of each randomized data is close to the result whole dataset (0.772 (randomized data) vs. 0.774 (whole Protegen dataset)). The randomized results of other methods are not consistent as Jenner-Predict. Standard deviation values again corroborate the consistent performance of Jenner-Predict (Table 2.6B).

Table 2.6B: Sensitivity of random datasets from vaccine candidate reported in Protegen database by software, NERVE, and web servers, Vaxign, VaxiJen and Jenner-Predict.

Sr no.	Dataset Details	NERVE	Vaxign	VaxiJen	Jenner-Predict
1	25%	0.622	0.422	0.511	0.733
2	25%	0.733	0.511	0.533	0.800
3	25%	0.577	0.444	0.488	0.733
4	25%	0.711	0.488	0.644	0.755
5	25%	0.600	0.444	0.577	0.800
Mean of 25% random data		***0.649***	***0.462***	***0.551***	***0.764***
6	50%	0.696	0.505	0.505	0.786
7	50%	0.696	0.494	0.550	0.797
8	50%	0.685	0.494	0.550	0.758
9	50%	0.617	0.426	0.606	0.797
10	50%	0.662	0.505	0.573	0.752
Mean of 50% random data		***0.671***	***0.485***	***0.557***	***0.778***
11	75%	0.669	0.503	0.533	0.789
12	75%	0.699	0.488	0.518	0.751
13	75%	0.699	0.518	0.526	0.774
14	75%	0.661	0.481	0.496	0.796
15	75%	0.654	0.458	0.556	0.759
Mean of 75% random data		***0.676***	***0.490***	***0.526***	***0.773***
Overall mean and standard deviation for three different random datasets*		**0.665 (0.014)**	**0.479 (0.015)**	**0.545 (0.016)**	**0.772 (0.007)**

* Standard deviation values are in parentheses.

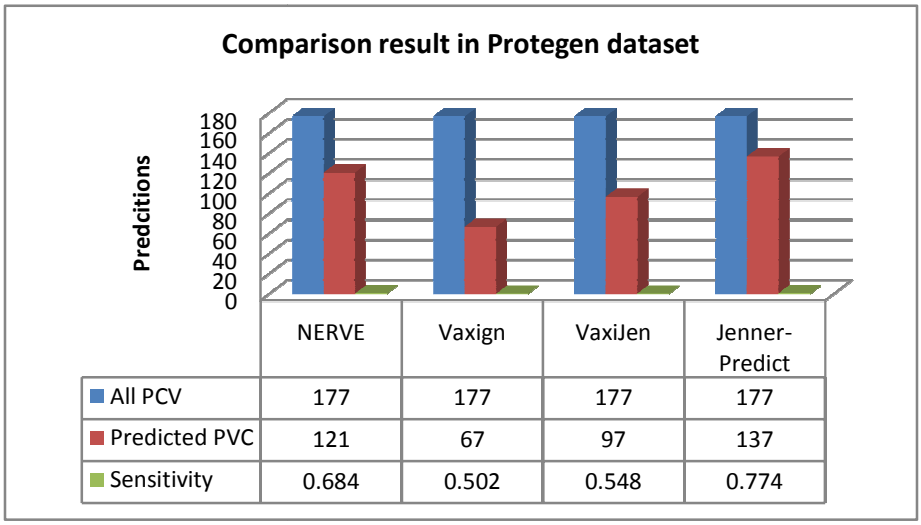


Figure 2.6: Comparative results of predicted PVCs in Protegen dataset through different methods.

Table 2.6A: Results of protein vaccine candidate (PVC) prediction from vaccine candidate reported in Protegen database by software, NERVE, and web servers, Vaxign, VaxiJen and Jenner-Predict*.

#S. No.	Protegen Database ID	Gene ID	Organism	$Gram +/-	NERVE	Vaxign	VaxiJen	Jenner-Predict
1.	VO_0011020	52630374	*Actinobacillus pleuropneumoniae*	N	YES	NO	NO	YES
2.	VO_0011022	190150285	*Actinobacillus pleuropneumoniae*	N	NO	NO	YES	YES
3.	VO_0010873	47566484	*Bacillus anthracis* str. 'Ames Ancestor	P	NO	NO	NO	YES
4.	VO_0010872	47566476	*Bacillus anthracis* str. 'Ames Ancestor'	P	NO	NO	NO	YES
5.	VO_0011030	3980256	*Bordetella pertussis*	N	YES	YES	YES	YES
6.	VO_0011031	580668	*Bordetella pertussis*	N	NO	NO	YES	YES
7.	VO_0011032	225311181	*Bordetella pertussis*	N	NO	NO	NO	YES

8.	VO_0011036	562026	*Bordetella pertussis*	N	YES	NO	YES	YES
9.	VO_0011037	225311180	*Bordetella pertussis*	N	YES	YES	NO	YES
10.	VO_0011038	225311183	*Bordetella pertussis*	N	YES	NO	NO	YES
11.	VO_0011039	225311182	*Bordetella pertussis*	N	YES	NO	NO	YES
12.	VO_0011033	33592195	*Bordetella pertussis* Tohama	N	NO	NO	YES	YES
13.	VO_0011034	33594638	*Bordetella pertussis* Tohama	N	NO	NO	NO	YES
14.	VO_0012384	11496927	*Borrelia burgdorferi* B31	N	YES	YES	YES	NO
15.	VO_0012385	11496910	*Borrelia burgdorferi* B31	N	NO	NO	YES	NO
16.	VO_0012386	11497024	*Borrelia burgdorferi* B31	N	YES	NO	YES	NO
17.	VO_0012365	62317941	*Brucella abortus*	N	NO	NO	YES	YES
18.	VO_0010948	122892474	*Brucella melitensis*	N	YES	NO	NO	YES
19.	VO_0010962	17986819	*Brucella melitensis* 16M	N	YES	NO	NO	NO
20.	VO_0010966	17987532	*Brucella melitensis* 16M	N	YES	YES	YES	YES
21.	VO_0010967	17987867	*Brucella melitensis* 16M	N	YES	YES	YES	YES
22.	VO_0010856	83269434	*Brucella melitensis* biovar Abortus	N	YES	NO	YES	NO
23.	VO_0010908	82700077	*Brucella melitensis* biovar Abortus	N	NO	NO	YES	NO
24.	VO_0010939	82700421	*Brucella melitensis* biovar Abortus	N	YES	YES	YES	YES

			2308					
25.	VO_0010971	82699574	*Brucella melitensis* biovar Abortus 2308	N	NO	NO	NO	YES
26.	VO_0010972	82700695	*Brucella melitensis* biovar Abortus 2308	N	YES	YES	YES	NO
27.	VO_0010973	82700483	*Brucella melitensis* biovar Abortus 2308	N	NO	NO	NO	YES
28.	VO_0011303	1929918	*Burkholderia pseudomallei*	N	YES	YES	YES	YES
29.	VO_0010922	53721504	*Burkholderia pseudomallei* K96243	N	NO	NO	NO	YES
30.	VO_0010956	1813949	*Campylobacter jejuni*	N	NO	NO	NO	YES
31.	VO_0011044	4704601	*Campylobacter jejuni*	N	YES	YES	YES	YES
32.	VO_0011047	116292649	*Campylobacter jejuni*	N	NO	YES	YES	NO
33.	VO_0011049	116292677	*Campylobacter jejuni*	N	YES	NO	YES	NO
34.	VO_0010955	57237809	*Campylobacter jejuni*	N	NO	NO	NO	YES
35.	VO_0011046	121612545	*Campylobacter jejuni*	N	YES	YES	YES	YES
36.	VO_0011048	121612344	*Campylobacter jejuni*	N	YES	YES	NO	YES
37.	VO_0010942	15792662	*Campylobacter jejuni*	N	YES	YES	YES	YES
38.	VO_0011045	112360246	*Campylobacter jejuni*	N	YES	NO	NO	YES
39.	VO_0010890	1518659	*Chlamydia muridarum*	N	YES	YES	YES	YES
40.	VO_0010885	15835130	*Chlamydia muridarum*	N	YES	NO	NO	YES
41.	VO_0010887	15835057	*Chlamydia*	N	NO	NO	NO	YES

			muridarum					
42.	VO_0010888	15835381	*Chlamydia muridarum*	N	NO	NO	NO	NO
43.	VO_0010889	15835382	*Chlamydia muridarum*	N	NO	NO	NO	NO
44.	VO_0010891	15834883	*Chlamydia muridarum*	N	YES	YES	NO	YES
45.	VO_0010893	15834882	*Chlamydia muridarum*	N	YES	YES	NO	YES
46.	VO_0010881	40601	*Chlamydophila abortus*	N	YES	YES	NO	YES
47.	VO_0011058	187438939	*Chlamydophila abortus*	N	YES	YES	YES	YES
48.	VO_0011057	62184917	*Chlamydophila abortus*	N	NO	NO	NO	NO
49.	VO_0011059	62184696	*Chlamydophila abortus*	N	NO	NO	NO	YES
50.	VO_0011060	62184824	*Chlamydophila abortus*	N	YES	NO	YES	NO
51.	VO_0010883	15618244	*Chlamydophila pneumoniae*	N	NO	YES	YES	YES
52.	VO_0010896	15618301	*Chlamydophila pneumoniae* CWL029	N	NO	NO	NO	NO
53.	VO_0010925	144545	*Chlamydophila psittaci*	N	YES	YES	NO	YES
54.	VO_0010900	241183337	*Clostridium botulinum*	P	YES	NO	NO	YES
55.	VO_0010904	169834607	*Clostridium botulinum*	P	NO	NO	NO	YES
56.	VO_0010905	217781	*Clostridium phage* c-st	P	YES	NO	NO	YES
57.	VO_0010909	157829735	Clostridium tetani	P	NO	NO	NO	YES
58.	VO_0011287	38199106	*Corynebacterium diphtheriae*	P	NO	NO	NO	YES
59.	VO_0010930	30025845	*Coxiella burnetii*	N	YES	YES	YES	NO
60.	VO_0010938	9632507	*Enterobacteria phage*		YES	NO	YES	YES

61.	VO_0010943	157021162	*Escherichia coli*	N	YES	YES	NO	YES
62.	VO_0010988	222104801	*Escherichia coli*	N	YES	YES	YES	YES
63.	VO_0010941	110643341	*Escherichia coli* 536	N	YES	YES	NO	YES
64.	VO_0010964	157418230	*Escherichia coli* APEC	N	YES	NO	YES	YES
65.	VO_0010965|	117624167	*Escherichia coli* APEC	N	YES	NO	YES	YES
66.	VO_0010984	26248334	*Escherichia coli* CFT073	N	YES	YES	YES	YES
67.	VO_0010985	26250982	*Escherichia coli* CFT073	N	NO	NO	YES	YES
68.	VO_0010986	26249458	*Escherichia coli* CFT073	N	YES	YES	YES	YES
69.	VO_0010989	26246291	*Escherichia coli* CFT073	N	YES	YES	YES	YES
70.	VO_0010990	26250246	*Escherichia coli*	N	YES	YES	YES	YES
71.	VO_0010940	15804222	*Escherichia coli* O157:H7 EDL933	N	YES	YES	YES	YES
72.	VO_0010944	15804220	*Escherichia coli* O157:H7 EDL933	N	YES	YES	YES	YES
73.	VO_0010993	209921909	*Escherichia coli* SE11	N	YES	YES	YES	YES
74.	VO_0010945	91213965	*Escherichia coli* UTI89	N	NO	NO	YES	YES
75.	VO_0011076	148688	*Francisella tularensis*	N	YES	YES	YES	NO
76.	VO_0011078	115315051	*Francisella tularensis* subsp. holarctica	N	YES	YES	YES	YES
77.	VO_0011070	118496734	*Francisella tularensis* subsp. novicida U112	N	YES	YES	NO	YES
78.	VO_0011072	56708413	*Francisella*	N	YES	NO	YES	NO

			tularensis subsp. tularensis SCHU					
79.	VO_0011077	56707247	*Francisella tularensis* subsp. tularensis SCHU	N	YES	NO	NO	NO
80.	VO_0010914	148896	*Haemophilus influenzae*	N	YES	YES	NO	NO
81.	VO_0010916	21686508	*Haemophilus influenzae*	N	YES	YES	YES	YES
82.	VO_0010917	23506944	*Haemophilus influenzae*	N	YES	YES	YES	YES
83.	VO_0010918	4574246	Haemophilus influenzae	N	NO	NO	YES	YES
84.	VO_0012376	2935168	*Haemophilus influenzae*	N	NO	NO	NO	YES
85.	VO_0011081	148971	*Haemophilus influenzae*	N	NO	NO	NO	YES
86.	VO_0012405	4929317	*Haemophilus influenzae*	N	YES	NO	NO	YES
87.	VO_0012406	9716645	*Haemophilus influenzae*	N	YES	YES	YES	YES
88.	VO_0010865	68249580	*Haemophilus influenzae*	N	YES	YES	YES	YES
89.	VO_0010915	68248984	*Haemophilus influenzae*	N	YES	YES	YES	YES
90.	VO_0011082	68249712	*Haemophilus influenzae*	N	YES	NO	YES	YES
91.	VO_0011083	68249503	*Haemophilus influenzae*	N	YES	YES	YES	YES
92.	VO_0011084	68248747	*Haemophilus influenzae*	N	YES	YES	YES	YES
93.	VO_0011339	3128145	*Helicobacter pylori*	N	NO	NO	NO	NO
94.	VO_0011341	50313203	*Helicobacter pylori*	N	YES	YES	YES	YES
95.	VO_0012388	16802248	*Listeria monocytogenes*	P	YES	NO	NO	YES

96.	VO_0012404	215422507	*Listeria monocytogenes*	P	YES	YES	YES	NO
97.	VO_0011344	26513901	*Listonella anguillarum*	N	YES	NO	YES	YES
98.	VO_0012366	224992215	*Mycobacterium bovis* BCG	P	YES	YES	NO	YES
99.	VO_0010936	29027587	*Mycobacterium tuberculosis*	N	YES	YES	NO	YES
100.	VO_0012373	148660242	*Mycobacterium tuberculosis* H37Ra	P	NO	NO	YES	NO
101.	VO_0010919	57117165	*Mycobacterium tuberculosis* H37Rv	P	YES	YES	0.5657	YES
102.	VO_0012364	15609023	*Mycobacterium tuberculosis* H37Rv	P	YES	YES	NO	YES
103.	VO_0010951	15611010	*Mycobacterium tuberculosis* H37Rv	P	YES	YES	YES	YES
104.	VO_0010953	15607267	*Mycobacterium tuberculosis* H37Rv	P	NO	NO	YES	YES
105.	VO_0012367	15609117	*Mycobacterium tuberculosis* H37Rv	P	YES	YES	NO	NO
106.	VO_0012370	57116798	*Mycobacterium tuberculosis* H37Rv	P	YES	YES	YES	YES
107.	VO_0012371	57116919	*Mycobacterium tuberculosis* H37Rv	P	YES	YES	NO	NO
108.	VO_0012372	57116920	*Mycobacterium tuberculosis* H37Rv	P	NO	NO	NO	YES
109.	VO_0011175	57116926	*Mycobacterium tuberculosis* H37Rv	P	YES	YES	NO	YES
110.	VO_0011176	15609063	*Mycobacterium tuberculosis* H37Rv	P	YES	NO	YES	YES
111.	VO_0012407	57116801	*Mycobacterium*	P	YES	YES	YES	YES

			tuberculosis H37Rv					
112.	VO_0012409	15610010	*Mycobacterium tuberculosis* H37Rv	P	NO	NO	NO	YES
113.	VO_0011178	7249262	*Mycoplasma gallisepticum*	N	YES	YES	NO	YES
114.	VO_0012380	8926211	*Neisseria meningitidis*	N	YES	YES	YES	YES
115.	VO_0011182	1017433	*Neisseria meningitidis*	N	YES	YES	YES	YES
116.	VO_0010946	15677945	*Neisseria meningitidis*	N	YES	NO	YES	YES
117.	VO_0010947	15677829	*Neisseria meningitidis*	N	NO	NO	NO	YES
118.	VO_0010974	15677776	*Neisseria meningitidis*	N	NO	NO	NO	YES
119.	VO_0010976	15677037	*Neisseria meningitidis*	N	YES	YES	NO	NO
120.	VO_0010977	15676883	*Neisseria meningitidis*	N	YES	YES	YES	YES
121.	VO_0010978	15675973	*Neisseria meningitidis*	N	YES	YES	NO	NO
122.	VO_0010979	15676561	*Neisseria meningitidis*	N	YES	YES	YES	YES
123.	VO_0010980	15677822	*Neisseria meningitidis*	N	NO	YES	YES	YES
124.	VO_0010981	15677705	*Neisseria meningitidis*	N	YES	YES	YES	NO
125.	VO_0010982	15677911	*Neisseria meningitidis*	N	YES	YES	NO	NO
126.	VO_0010983	15676917	*Neisseria meningitidis*	N	YES	YES	YES	NO
127.	VO_0011180	15676020	*Neisseria meningitidis*	N	YES	YES	YES	YES
128.	VO_0012360	15596974	*Pseudomonas aeruginosa*	N	YES	YES	YES	YES
129.	VO_0012361	15596903	*Pseudomonas aeruginosa* PAO1	N	NO	NO	NO	YES

130.	VO_0012362	15598049	*Pseudomonas aeruginosa* PAO1	N	YES	NO	YES	NO
131.	VO_0011317	152498	*Rickettsia prowazekii*	N	YES	YES	YES	YES
132.	VO_0011234	112710	*Rickettsia rickettsii*	N	YES	YES	YES	YES
133.	VO_0011235	6685726	*Rickettsia rickettsii*	N	YES	YES	YES	YES
134.	VO_0010997	259475459	*Salmonella enterica* subsp. enterica serovar Typhimurium	N	YES	YES	YES	YES
135.	VO_0010999	54036439	*Salmonella enterica* subsp. enterica serovar Typhimurium	N	NO	NO	NO	YES
136.	VO_0010994	22036246	*Shigella flexneri* 2a	N	YES	YES	NO	YES
137.	VO_0010995	47056	*Shigella flexneri* 2a	N	YES	NO	YES	YES
138.	VO_0010996	22036352	*Shigella flexneri* 2a	N	YES	NO	NO	YES
139.	VO_0010992	13449098	*Shigella flexneri* 5a	N	YES	NO	YES	NO
140.	VO_0012392	120457	*Staphylococcus aureus*	P	YES	YES	YES	YES
141.	VO_0012391	49482291	*Staphylococcus aureus* subsp. aureus MRSA252	P	NO	NO	NO	YES
142.	VO_0012390	151220968	*Staphylococcus aureus* subsp. aureus	P	YES	NO	YES	YES
143.	VO_0012393	5327234	*Streptococcus agalactiae*	P	YES	YES	YES	YES
144.	VO_0012403	1620648	*Streptococcus agalactiae*	P	YES	NO	YES	YES
145.	VO_0012402	76788047	*Streptococcus agalactiae*	P	YES	YES	YES	YES

146.	VO_0011211	225870123	*Streptococcus equi*	P	YES	YES	NO	YES
147.	VO_0011213	225870316	*Streptococcus equi*	P	YES	YES	YES	YES
148.	VO_0011215	225869898	*Streptococcus equi*	P	YES	NO	YES	NO
149.	VO_0011216	225871286	*Streptococcus equi*	P	NO	NO	NO	YES
150.	VO_0011214	225869227	*Streptococcus equi* subsp. zooepidemicus	P	YES	YES	YES	YES
151.	VO_0011190	209867628	*Streptococcus pneumoniae*	P	NO	NO	NO	YES
152.	VO_0011192	116515376	*Streptococcus pneumoniae* D39	P	YES	NO	NO	YES
153.	VO_0011203	116515359	*Streptococcus pneumoniae* D39	P	YES	YES	YES	YES
154.	VO_0011204	116515876	*Streptococcus pneumoniae* D39	P	NO	YES	NO	YES
155.	VO_0012396	14915682	*Streptococcus pyogenes*	P	YES	NO	YES	YES
156.	VO_0012398	90567992	*Streptococcus pyogenes* serotype M12	P	YES	YES	NO	NO
157.	VO_0010901	166236883	synthetic construct	P	YES	YES	NO	YES
158.	VO_0011173	283801990	synthetic construct	S	YES	YES	YES	YES
159.	VO_0012399	15639249	*Treponema pallidum*	N	NO	NO	NO	YES
160.	VO_0012400	15639756	*Treponema pallidum*	N	NO	NO	YES	NO
161.	VO_0012401	159158965	*Treponema pallidum*	N	YES	YES	YES	YES
162.	VO_0011270	110264635	*Vibrio cholerae* O1	N	NO	NO	YES	YES
163.	VO_0011271	21616882	*Vibrio cholerae* O1	N	YES	YES	NO	YES

164.	VO_0011272	15640854	*Vibrio cholerae* O1 biovar El Tor str. N16961	N	YES	YES	YES	YES
165.	VO_0011003	162417777	*Yersinia pestis* Angola	N	NO	NO	NO	NO
166.	VO_0010870	45478667	*Yersinia pestis* biovar Microtus str. 91001]	N	YES	YES	YES	NO
167.	VO_0010874	16082719	*Yersinia pestis* CO92	N	NO	NO	NO	NO
168.	VO_0010877	16082686	*Yersinia pestis* CO92	N	YES	YES	YES	NO
169.	VO_0010878	16082716	*Yersinia pestis* CO92	N	NO	NO	NO	YES
170.	VO_0012359	16082755	*Yersinia pestis* CO92	N	NO	NO	YES	YES
171.	VO_0010913	16082743	*Yersinia pestis* CO92	N	NO	NO	NO	YES
172.	VO_0011005	218927800	*Yersinia pestis* CO92	N	YES	YES	YES	YES
173.	VO_0011007	218927806	*Yersinia pestis* CO92	N	NO	NO	NO	YES
174.	VO_0011009	218930092	*Yersinia pestis* CO92	N	NO	NO	NO	NO
175.	VO_0011010	218928525	*Yersinia pestis* CO92	N	YES	NO	NO	NO
176.	VO_0011012	218927621	*Yersinia pestis* CO92	N	NO	NO	YES	NO
177.	VO_0011013	218930726	*Yersinia pestis* CO92	N	NO	NO	NO	YES

* For details, see methods section. Jenner-Predict server is based on domains involved in host-pathogen interactions which are important in pathogenesis and disease establishment. Out of total 257 bacterial protective PVCs reported in the Protegen database, 177 bacterial protective antigens having less than 90 percent identity were selected for evaluation purpose from 200 proteins with non-cytosolic cellular localization and having less than two transmembrane helices. For comparison with VaxiJen, a cut-off of 0.6 was used instead of default parameter 0.4 as it predicts almost half of proteome as vaccine candidates with default parameter.

S. No. indicates Serial Number; p or n in Gram column indicates gram positive and gram negative, respectively; and YES or NO denotes the corresponding protein is predicted or not, respectively by the corresponding software or web server.

$ Gram positive bacteria are denoted as "P" and gram negative bacteria are denoted as "N" in the above table.

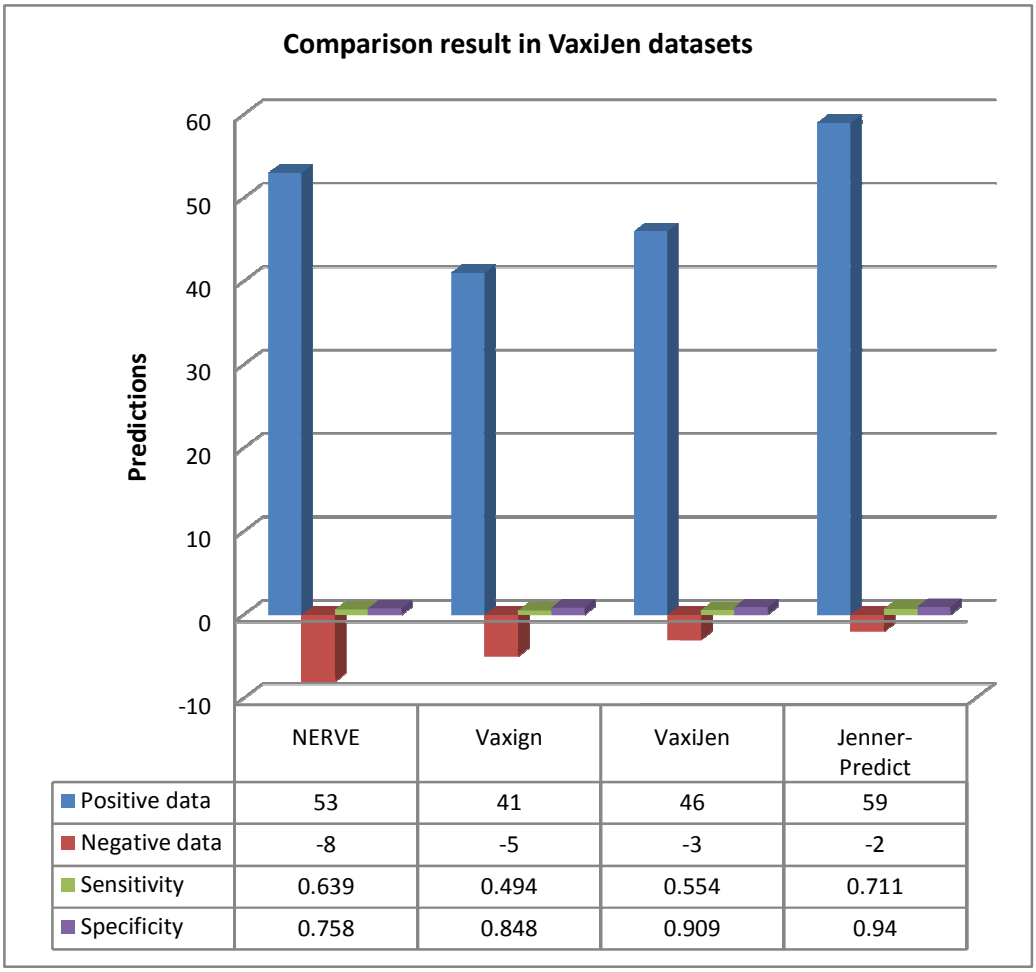


Figure 2.7: Comparative result of predicted PVCs in VaxiJen datasets through different methods

Table 2.7: Results of protein vaccine candidate (PVC) prediction from positive dataset used for VaxiJen server development by software, NERVE, and web servers, Vaxign, VaxiJen and Jenner-Predict*.

S. No.	SwissProt ID	Gram (P/N)	Localization	Organism	Nerve	VaxiJen	Vaxign	Jenner-Predict
1.	Q9RN24	P	Extracellular	*Bacillus anthracis*	YES	YES	YES	YES
2.	Q93V22 P04977	N	Extracellular	*Bordetella pertussis*	NO	NO	NO	YES
3.	Q9S6N1	N	Outer Membrane	*Bordetella pertussis*	NO	YES	NO	YES
4.	P14013	N		*Borrelia burgdorferi*	YES	YES	NO	NO

5.	P17739	N	Outer Membrane	*Borrelia burgdorferi*	YES	YES	YES	NO
6.	P70854	N	Unknown	*Borrelia burgdorferi*	NO	NO	NO	YES
7.	Q07337	N	Outer Membrane	*Borrelia burgdorferi*	YES	YES	NO	NO
8.	P0A470	N	Unknown	*Brucella abortus*	NO	NO	NO	NO
9.	P15453	N	Periplasmic	*Brucella abortus*	YES	YES	NO	NO
10.	Q45321	N	Outer Membrane	*Brucella melitensis*	YES	YES	YES	YES
11.	P27053	N	Extracellular	*Campylobacter coli*	YES	YES	YES	YES
12.	Q46412	N	Outer Membrane	*Chlamydia trachomatis*	YES	YES	YES	YES
13.	Q9RF12	P	Extracellular	*Clostridium perfringens*	YES	YES	YES	NO
14.	Q9RM68	P	Unknown	*Clostridium perfringens*	YES	YES	NO	YES
15.	Q9LA13	P	Extracellular	*Clostridium tetani*	NO	NO	NO	YES
16.	P20626	P	Unknown	*Corynebacterium pseudotuberculosis*	YES	NO	YES	NO
17.	P05825	N	Outer Membrane	*Escherichia coli*	YES	YES	YES	YES
18.	P08191	N	Unknown	*Escherichia coli*	YES	YES	YES	YES
19.	P0AFZ6	N	Cytoplasmic Membrane	*Escherichia coli*	NO	NO	NO	NO
20.	P17315	N	Outer Membrane	*Escherichia coli*	YES	YES	NO	YES
21.	Q93V32	N	Unknown	*Escherichia coli*	YES	NO	YES	YES
22.	P43838	N	Outer Membrane	*Haemophilus influenzae*	YES	YES	YES	YES
23.	P10324	N	Outer Membrane	*Haemophilus influenzae*	YES	YES	YES	YES
24.	P45996	N	Outer Membrane	*Haemophilus influenzae*	YES	YES	NO	YES
25.	Q9ZKX5	N	Unknown	*Helicobacter pylori*	NO	NO	NO	NO
26.	P24017	N	Outer Membrane	*Klebsiella pneumoniae*	YES	YES	NO	YES
27.	Q48427	N	Outer Membrane	*Klebsiella pneumoniae*	YES	YES	YES	YES
28.	Q48473	N	Outer Membrane	*Klebsiella pneumoniae*	YES	YES	NO	YES
29.	P21347	N	Extracellular	*Klebsiella pneumoniae*	YES	YES	YES	NO
30.	Q9Z374	N	Unknown	*Klebsiella*	YES	NO	YES	NO

				pneumoniae				
31.	P21171	P	Extracellular	*Listeria monocytogenes*	YES	YES	YES	YES
32.	Q9L5B9	P	Extracellular	*Listeria monocytogenes*	YES	NO	NO	YES
33.	Q06947	P	Extracellular	*Mycobacterium avium*	YES	YES	YES	YES
34.	P0A4V3 P0C926	N	Outermembrane	*Mycobacterium bovis*	YES	YES	NO	NO
35.	P0A671	P	Cytoplasmic Membrane	*Mycobacterium bovis*	NO	NO	NO	YES
36.	O05870	P	Extracellular	*Mycobacterium tuberculosis*	YES	YES	YES	YES
37.	P0A564	P	Extracellular	*Mycobacterium tuberculosis*	YES	YES	YES	YES
38.	P0A5P6	P	Cell wall	*Mycobacterium tuberculosis*	NO	NO	YES	NO
39.	P0A5Q2	P	Extracellular	*Mycobacterium tuberculosis*	YES	YES	NO	YES
40.	P0A5Q4	P	Extracellular	*Mycobacterium tuberculosis*	YES	YES	YES	NO
41.	P0A5Y2	P	Unknown	*Mycobacterium tuberculosis*	YES	YES	YES	YES
42.	P0A670	P	Cytoplasmic Membrane	*Mycobacterium tuberculosis*	NO	NO	YES	YES
43.	P31952 A5U3Q3	P	Extracellular	*Mycobacterium tuberculosis*	YES	YES	NO	YES
44.	Q79F92	P	Cytoplasmic Membrane	*Mycobacterium tuberculosis*	NO	NO	NO	NO
45.	O07175	P	Unknown	*Mycobacterium tuberculosis*	NO	YES	NO	YES
46.	O50430	P	Cytoplasmic Membrane	*Mycobacterium tuberculosis*	NO	NO	YES	NO
47.	P0A4V6	P	Extracellular	*Mycobacterium tuberculosis*	YES	YES	YES	YES
48.	P0A566	P	Extracellular	*Mycobacterium tuberculosis*	YES	YES		YES
49.	P0A568	P	Unknown	*Mycobacterium tuberculosis*	YES	YES	YES	YES
50.	P0A5B7	P	Cell wall	*Mycobacterium*	YES	YES	NO	NO

				tuberculosis				
51.	P0A5P2	P	Extracellular	*Mycobacterium tuberculosis*	NO	YES	NO	NO
52.	P0A5P8	P	Extracellular	*Mycobacterium tuberculosis*	YES	YES	YES	NO
53.	P0A5Y2	P	Unknown	*Mycobacterium tuberculosis*	YES	YES	YES	YES
54.	P15712	P	Unknown	*Mycobacterium tuberculosis*	YES	YES	YES	YES
55.	P65306	P	Cytoplasmic Membrane	*Mycobacterium tuberculosis*	NO	NO	NO	NO
56.	Q7D8M9	P	Cytoplasmic Membrane	*Mycobacterium tuberculosis*	NO	NO	YES	YES
57.	P96943	N	Outer Membrane	*Neisseria meningitidis*	YES	YES	YES	YES
58.	Q53348	N	Outer Membrane	*Neisseria meningitidis*	YES	YES	YES	YES
59.	Q53990	N	Extracellular	*Neisseria meningitidis*	YES	YES	NO	YES
60.	O30527	N	Extracellular	*Pseudomonas aeruginosa*	NO	NO	NO	YES
61.	P11439	N	Extracellular	*Pseudomonas aeruginosa*	NO	NO	NO	YES
62.	32722	N	Outer Membrane	*Pseudomonas aeruginosa*	NO	YES	YES	YES
63.	P13794	N	Outer Membrane	*Pseudomonas aeruginosa*	YES	YES	YES	YES
64.	Q8ZP50	N	Outer Membrane	*Salmonella typhimurium*	YES	YES	YES	YES
65.	P69178	N	Unknown	*Shigella dysenteriae*	YES	YES	NO	YES
66.	P0A0L2	P	Extracellular	*Staphylococcus aureus*	YES	YES	NO	YES
67.	Q53653	P	Cell wall	*Staphylococcus aureus*	YES	YES	NO	YES
68.	Q3K3Z5	P	Extracellular	*Streptococcus agalactiae*	YES	YES	YES	YES
69.	Q9ZHG7	P	Unknown	*Streptococcus agalactiae*	NO	NO	NO	NO
70.	O34097	P	Extracellular	*Streptococcus pneumoniae*	NO	YES	YES	YES

71.	P11990 P0C2J9	P	Extracellular	*Streptococcus pneumoniae*	YES	NO	NO	YES
72.	Q8DN05	P	Extracellular	*Streptococcus pneumoniae*	YES	YES	YES	YES
73.	Q8VQ82	P	Cytoplasmic Membrane	*Streptococcus pneumoniae*	NO	NO	YES	YES
74.	Q9AG74	P	Unknown	*Streptococcus pneumoniae*	NO	NO	NO	NO
75.	P59206	P	Extracellular	*Streptococcus pneumoniae*	YES	NO	YES	YES
76.	Q9Z4J8	P	Unknown	*Streptococcus pneumoniae*	YES	NO	YES	YES
77.	O30405	N	Outer Membrane	*Treponema pallidum*	NO	NO	NO	YES
78.	O83867	N	Unknown	*Treponema pallidum*	NO	YES	NO	YES
79.	P19649	N	Outer Membrane	*Treponema pallidum*	NO	YES	NO	NO
80.	Q87L97	N	Cytoplasmic Membrane	*Vibrio parahaemolyticus*	NO	NO	NO	NO
81.	P21206 A4TSQ1	N	Extracellular	*Yersinia pestis*	NO	NO	NO	YES
82.	P26948	N	Extracellular	*Yersinia pestis*	YES	YES	YES	NO
83.	Q7DHH4	P	Cytoplasmic Membrane	*Staphylococcus aureus*	NO	NO	NO	YES

* See details in methods section. Jenner-Predict server has been developed by us and is based on domains involved in host-pathogen interactions which are important in pathogenesis and disease establishment. Out of total 100 bacterial protective antigen in positive dataset used for VaxiJen server development, 83 proteins with non-cytosolic localization and having less than two transmembrane helices were selected for evaluation by different methods. For VaxiJen, a cut-off of 0.6 was used instead of default parameter 0.4 as it predicts almost half of a bacterial proteome as vaccine candidates with default parameter.

S. No. indicates Serial Number; P or N in Gram column indicate gram positive or gram negative, respectively; and YES or NO denotes the corresponding protein is predicted or not, respectively by the corresponding software or web server.

Table 2.8: Results of protein vaccine candidate (PVC) prediction from negative dataset used for VaxiJen server development by software, NERVE, and web servers, Vaxign, VaxiJen and Jenner-Predict*

#S. No.	SwissProt ID	Organism	Gram	Nerve	VaxiJen	Vaxign	Jenner-Predict
1.	P26826	*Clostridium perfringens*	P	YES	NOT	YES	NOT
2.	Q8XMI8	*Clostridium perfringens*	P	NOT	NOT	NOT	NOT
3.	Q890Y8	*Clostridium tetani*	P	NOT	NOT	NOT	NOT
4.	Q8Y652	*Listeria monocytogenes*	P	NOT	NOT	NOT	NOT
5.	Q48909	*Mycobacterium avium*	P	NOT	NOT	NOT	NOT
6.	Q48919	*Mycobacterium avium*	P	YES	NOT	YES	YES
7.	Q7TXM7	*Mycobacterium bovis*	P	NOT	YES	NOT	NOT
8.	O69742	*Mycobacterium tuberculosis*	P	YES	YES	YES	NOT
9.	O69743	*Mycobacterium tuberculosis*	P	NOT	NOT	NOT	NOT
10.	P0A5R6	*Mycobacterium tuberculosis*	P	NOT	NOT	NOT	NOT
11.	Q79F93	*Mycobacterium tuberculosis*	P	YES	NOT	NOT	NOT
12.	P0A5N0	*Mycobacterium tuberculosis*	P	NOT	NOT	NOT	NOT
13.	P63338	*Mycobacterium tuberculosis*	P	NOT	NOT	NOT	NOT
14.	P64249	*Mycobacterium tuberculosis*	P	NOT	NOT	NOT	NOT
15.	P96910	*Mycobacterium tuberculosis*	P	NOT	NOT	NOT	NOT
16.	O07341	*Streptococcus pneumoniae*	P	NOT	NOT	NOT	NOT
17.	O33754	*Streptococcus pneumoniae*	P	NOT	NOT	NOT	NOT
18.	Q7VWV9	*Bordetella pertussis*	N	NOT	NOT	NOT	NOT
19.	O51043	*Borrelia burgdorferi*	N	NOT	NOT	NOT	NOT
20.	O51240	*Borrelia burgdorferi*	N	NOT	NOT	NOT	NOT
21.	Q05051	*Borrelia burgdorferi*	N	YES	NOT	YES	NOT
22.	P0AD79	*Escherichia coli*	N	NOT	NOT	NOT	NOT
23.	O25798	*Helicobacter pylori*	N	YES	NOT	NOT	NOT
24.	Q3S3S0	*Helicobacter pylori*	N	NOT	NOT	NOT	NOT

25.	P20440	*Klebsiella pneumoniae*	N	YES	NOT	NOT	NOT
26.	Q48439	*Klebsiella pneumoniae*	N	NOT	NOT	NOT	NOT
27.	Q84HD6	*Neisseria meningitidis*	N	NOT	NOT	NOT	NOT
28.	P23181	*Pseudomonas aeruginosa*	N	NOT	NOT	NOT	NOT
29.	Q53247	*Orientia tsutsugamushi*	N	NOT	NOT	NOT	YES
30.	P07643	*reponema pallidum*	N	NOT	YES	NOT	NOT
31.	P29724	*Treponema pallidum*	N	NOT	NOT	NOT	NOT
32.	Q8ZF61	*Yersinia pestis*	N	YES	NOT	YES	NOT
33.	Q8ZIC6	*Yersinia pestis*	N	NOT	NOT	NOT	NOT

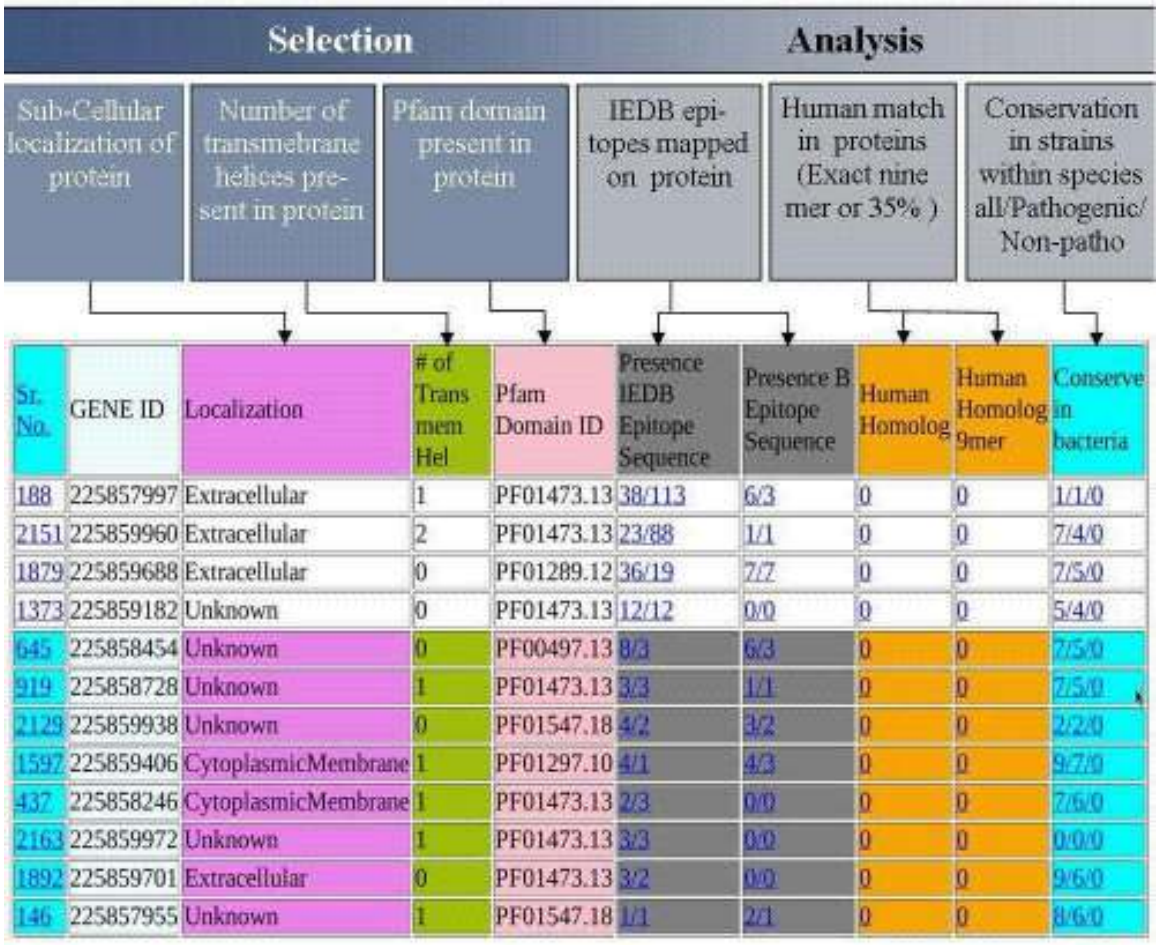
* See details in methods section. Jenner-Predict server is based on domains involved in host-pathogen interactions which are important in pathogenesis and disease establishment. Out of total 100 bacterial protective antigen in negative dataset used for VaxiJen server, 34 proteins with non-cytosolic cellular localization and having less than two transmembrane helices were selected evaluation by different methods. For VaxiJen, a cut-off of 0.6 was used instead of default parameter 0.4 as it predicts almost half of a bacterial proteome as vaccine candidates with default parameter.

S. No. indicates Serial Number; P or N in Gram column indicate gram-positive or gram-negative, respectively; and YES or NO denotes the corresponding protein is predicted or not, respectively by the corresponding software or web server.

2.3.4 ***Output***

The server, Jenner-Predict, has been designed for easy submission of a job in three different ways as well as user-friendly and interactive interpretation of results in the form of table and hyperlinks on output values. Just after job submission, a unique URL link is generated through random numbers so that users actively remain confident. The user may bookmark the unique URL link of his/her submission for tracking the status as jobs are processed in a queue. Once a job has been completed, the output is provided in a tabular format which can be accessed through unique URL link at any time. A sample output of result is represented in Figure 2.8. The information provided in different columns are as follows: 1. Sr. No.; 2. Gene Id; 3. Localization; 4. No. of transmembrane helices; 5. Pfam domain ID; 6. No. of IEDB TCE(s) match(s); 7. No. of IEDB BCE(s) match(s) (Hyperlinks on 6 and 7 showing details of matching epitopes); 8. and 9. Autoimmunity information through 35% identical matches in 80 AA lengths, and No. of continuous 9-mer identical

match in an alignment, respectively; and 10. Conservation in number of strains of an organism in the form of x/y/z: x. all (pathogenic and non-pathogenic)/, y. pathogenic/, z. non-pathogenic.



Sr. No.	GENE ID	Localization	# of Trans mem Hel	Pfam Domain ID	Presence IEDB Epitope Sequence	Presence B Epitope Sequence	Human Homolog	Human Homolog 9mer	Conserve in bacteria
188	225857997	Extracellular	1	PF01473.13	38/113	6/3	0	0	1/1/0
2151	225859960	Extracellular	2	PF01473.13	23/88	1/1	0	0	7/4/0
1879	225859688	Extracellular	0	PF01289.12	36/19	7/7	0	0	7/5/0
1373	225859182	Unknown	0	PF01473.13	12/12	0/0	0	0	5/4/0
645	225858454	Unknown	0	PF00497.13	8/3	6/3	0	0	7/5/0
919	225858728	Unknown	1	PF01473.13	3/3	1/1	0	0	7/5/0
2129	225859938	Unknown	0	PF01547.18	4/2	3/2	0	0	2/2/0
1597	225859406	CytoplasmicMembrane	1	PF01297.10	4/1	4/3	0	0	9/7/0
437	225858246	CytoplasmicMembrane	1	PF01473.13	2/3	0/0	0	0	7/6/0
2163	225859972	Unknown	1	PF01473.13	3/3	0/0	0	0	0/0/0
1892	225859701	Extracellular	0	PF01473.13	3/2	0/0	0	0	9/6/0
146	225857955	Unknown	1	PF01547.18	1/1	2/1	0	0	8/6/0

Figure 2.8: Output result of Jenner-Predict server

2.4 Discussion

The motivation behind developing this web server is to provide credible vaccine candidates and information regarding their vaccine potential in terms of possible immunogenicity, absence of autoimmunity and conservation so that subunit vaccine development can be accelerated. The outcome of the web server has substantiated that domains involved in host-pathogen interactions are better criterion for prediction of PVCs than approaches dependent upon only adhesin-likeliness or machine learning. As PVCs are predicted based on their functions, biologists can assess the importance of given function in

pathogenesis for that organism. The information regarding the function of PVC could be instrumental for vaccine development. For example, colonization is crucial in *Streptococcus* pathogenesis and proteins (also predicted by Jenner-Predict) involved in this process were used as vaccine candidates (Nobbs *et al*., 2009).

Most of the earlier RV methods focused on outer membrane or secretory proteins of a proteome to identify PVCs. Pizza *et al.* screened proteome sequences of *N. meningitis* to identify proteins which are probably surface exposed or involved in transportation and obtained 570 proteins. Out of them, 350 proteins were expressed and experimentally tested for their immunogenic potential. Finally, 7 proteins were found to provide protective immunity against *N. meningitidis* (Pizza *et al*., 2000). Similarly, Wizemann *et al.* searched for motifs related to secretory or surface binding proteins in *S. pneumoniae* proteome and 130 proteins were identified. Out of them, 108 were expressed and tested for their protective immunity. Finally, 6 proteins were found as protective antigens. Similar studies were performed for *P. gingivalis* (Ross *et al*., 2001) and *C. pneumoniae* (Montigiani *et al*., 2002). Although proteins providing immunity were identified in all the above mentioned studies but the number of experiments, cost and time requirement were enormous even for identifying PVCs from a particular localization. On the contrary, Jenner-Predict server is relying on protein domains involved in host-pathogen interactions for providing reasonably less number of prioritized vaccine candidates from a proteome. For better validation of vaccine candidates, the user may select few prospective vaccine candidates for experimental testing to verify their protective immunity.

The BCE and TCE mapping algorithms were developed to identify possible immunogenic region(s) and consequently prediction of immunogenic potential of a protein. But these methods have drawbacks of over-prediction and even predict epitope(s) in known non-antigenic proteins (Blythe and Flower, 2005; Gowthaman and Agrewala, 2007; Ponomarenko and Bourne, 2007; Zhang *et al*., 2010). Currently available antigen or PVC prediction methods were not validated on complete or diverse data. NERVE software was evaluated on its prediction ability of popular vaccine candidates from five bacterial proteomes instead of all known vaccine candidates in those organisms. Similarly, the Vaxign server was evaluated against only limited number of known OMP vaccine candidates from uropathogenic *E. coli* (He *et al*., 2010). Further, the VaxiJen server was developed on limited data. Even some of the sequences used as the negative data (non-antigenic) for web server development were predicted as vaccine candidates (antigenic proteins) by other methods (Table 2.8). Our web

server predicted two such proteins (Q48919 and Q53247) as PVCs from the negative dataset sequences. Experimental data confirmed that alanine and proline rich secreted protein (Q48919) is immunogenic (Lee *et al*., 2009) whereas the other protein is a periplasmic serine or membrane protease (htrA gene) and has already been reported as protective antigen in *Haemophilus influenzae* (Loosmore *et al*., 1998). The other PVCs predicted from negative dataset by NERVE, Vaxign and VaxiJen do not have evidence of being immunogenic in literature. This outcome justifies higher sensitivity of our method.

To provide prospective PVCs of a proteome, the predicted vaccine candidates are prioritized by Jenner-Predict server. The PVCs having more IEDB epitope matches are ranked higher as such epitope match increases their possibility to be immunogenic. Since epitopes identified using 'hands on' peptide-by-peptide in vitro assays have been more substantive than epitopes predicted by using *in silico* methods, known and validated epitopes from the IEDB (Vita *et al*., 2010) are mapped on PVCs to predict their potential immunogenic regions. The web server de-prioritizes PVC having human homolog(s) as they can potentially cause autoimmunity (Iwai *et al*., 2005) or produce low immune response (Grossman and Paul, 2001). Conservation information of PVCs is provided to demonstrate their broad specificities. Since the web server provides conserved and potential immunogenic PVCs, it may be useful to replace the existing strain-specific vaccine candidates. For example, the established vaccine candidate, PspA, is having choline-binding protein (CBP) domain and it has limited application from vaccine point of view as it is strain-specific. In contrast, our tool predicted cbpE (GI: 225858728) protein which is conserved across different strains of *S. pneumoniae* and has the same CBP domain. Since this protein has been surface exposed, and involved in nasopharyngeal colonization and/or dissemination of *S. pneumoniae* which is important for virulence, this protein may further be explored for vaccine development process (Rosenow *et al*., 1997).

Jenner-Predict server predicted 3 experimentally known promising PVCs, STK (Giefing *et al*., 2008), NanA (Turbyfill *et al*., 2008), and PsaA (Talkington *et al*., 1996) in *S. pneumoniae* which are containing domains from non-adhesins functional classes such as PASTA, BNR and SBP, respectively, and these proteins are known to be immunogenic. Similarly, our method predicted established non-adhesin vaccine candidates, Omp A, IreA, BtuB and TolC (provides protective immune responses (Hagan and Mobley, 2007) in *E. coli*). The wide-ranging applicability of the web server to all bacteria is substantiated by its high sensitivity for predicting diverse protective antigens from more than 40 pathogenic bacteria reported in

Protegen database (Table 2.6) and dataset used for VaxiJen server development (Table 2.7 and Table 2.8). Our domain based method was effective in predicting many established non-adhesin vaccine candidates reported in Protegen database (Yang *et al*., 2011) such as 13 toxins, 12 binding proteins (fibronectin, penicillin, choline, etc.), 10 membrane proteins, 6 surface proteins, etc. (Table 2.6) which were not predicted by other methods. These protective antigens are involved in many important pathogenesis processes like virulence, invasion, colonization, iron acquisition, osmo-regulation, etc. (Table 2.6). Additional significance of comparison result was also evaluated on fifteen random datasets where in all cases Jenner-Predict outperform other methods (Table 2.6b). Furthermore mean of sensitivity again justify the predictive dominance of Jenner-Predict and minimum standard deviation of sensitivity within random datasets supports sensitive of method followed in Jenner-Predict (Table 2.6b). The sensitivity of the web server was further substantiated by its prediction of immunogenic protein, Q48919, from negative dataset used for training of VaxiJen server (Table 2.8). Higher sensitivity and specificity of Jenner-Predict server (Figure 2.7) justifies the domains involved in host-pathogen interactions and pathogenesis are better criteria for PVCs prediction than other existing approaches. Jenner-Predict server is freely accessible for research and education purpose at http://14.139.240.55/vaccine/home.html. Output results and associated data can also be downloaded from available links. Standalone version of tool is not available.

Further Jenner-Predict can be improved by including prediction of vaccine candidates from cytosolic proteins. The methodology followed in Jenner-Predict (currently considers non-cytosolic proteins), domain based approach, may extended further for vaccine candidate prediction. Considering cytosolic protein on the basis of any sequence property or using any algorithm may further improve the prediction accuracy of the method.

2.5 Conclusions

The Jenner-Predict server has been developed to predict potential PVCs and also to provide their vaccine potential with an objective of assisting subunit vaccine development. The web server was validated on independent and diverse datasets, where it outperformed other PVC prediction tools. Its performance substantiated that the proteins involved in host-pathogen interactions and pathogenesis are better criteria than methods based on machine learning or adhesin-likeliness. Our method predicts less number of proteins with high prediction accuracy which confirms its reliability. Mapping of known epitopes from IEDB

database on PVCs increases the probability of a protein to be immunogenic. Comparison of these PVCs with human proteome sequences reduces the chance of their failure due to autoimmunity. Conservation of PVCs in pathogenic strains provides crucial information on their broad-specificities or sero-independent nature. The web server demonstrated that domain-based method can be used to predict PVCs from pathogen proteomes. Since the web server provides prioritized PVCs, few prospective proteins from a proteome could be taken for experimental evaluation to identify subunit vaccine candidates.

CHAPTER 3

EPICOMBFLU: EXPLORING KNOWN INFLUENZA EPITOPES AND THEIR COMBINATION TO DESIGN UNIVERSAL INFLUENZA VACCINE

ABSTRACT

In spite of concerted efforts by Global Influenza Surveillance Network (GISN) and World Health Organization (WHO), influenza is responsible for annual deaths of half a million people worldwide. Vaccination is the best preventive measure against this pervasive health problem but influenza vaccines developed from surveillance data of each season are strain-specific and provide protection against strains which are closely related to strains used in vaccine itself; therefore, these vaccines are unable to provide protection against pandemic strains arising from antigenic shift and/or drift. Seasonal epidemics and occasional pandemics along with drug resistance of influenza have created a necessity of universal influenza vaccine (UIV). Researchers have shown that combination of conserved epitopes has potential to be used as UIV. Influenza virus is one of the highly studied viruses due to its ancient importance. Hence, enormous information related to strains, proteins and nucleotides sequences, epitopes and other associated information are available for influenza virus. But there is no resource where conservation of epitopes is available as well as a resource where manually or automatically combination of conserved epitopes can be checked for their vaccine potential. In this chapter of thesis, available data on strains, proteins, epitopes and their associated information were used to develop a web-resource, EpiCombFlu which can explore different influenza epitopes and their combinations for conservation among different strains, population coverage, and immune response for vaccine design within the module "Epitope Information Resource". Forward selection algorithm (FSA) was implemented in EpiCombFlu to select optimum combination of epitopes (in module "Epitope Combination Explorer") which may be expressed and evaluated as potential UIV. The web-resource is freely available at http://14.139.240.55/influenza/home.html.

(The content of this chapter has already been published: V. Jaiswal, *et al.* **"EpiCombFlu: Exploring known influenza epitopes and their combination to design universal influenza vaccine"**. *Bioinformatics* 2013, **29**(15):1904-1907)

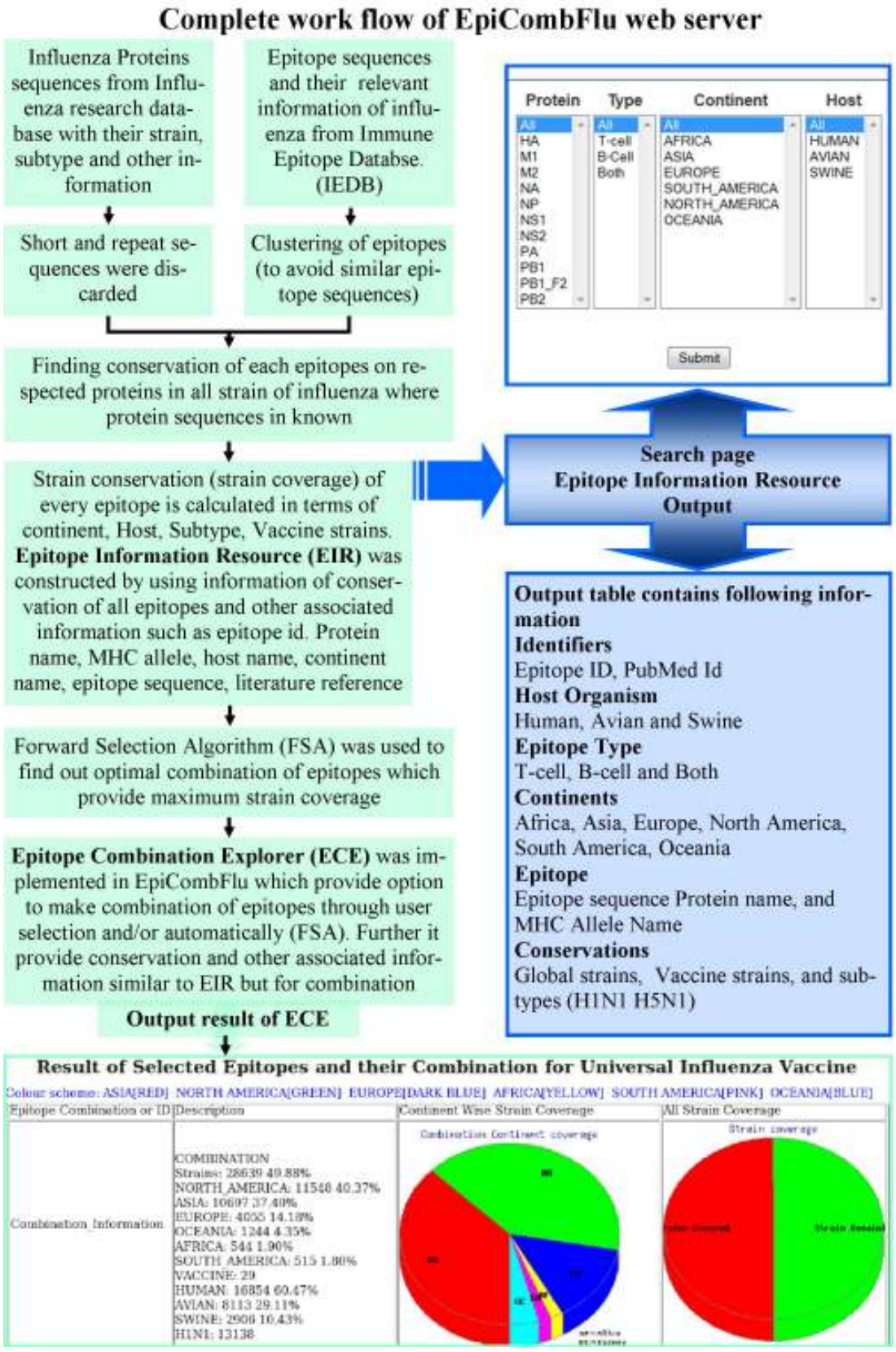


Figure 3.1: Graphical abstract of EpiCombFlu web server

3.1 Introduction

Old battle against influenza worldwide costs half a million lives annually. Global influenza epidemics causes massive burden on health care system with about 1 billion infections annually which resulted in around 5 million severe cases of disease (Girard *et al*., 2005). Despite availability of vaccines, there is always a threat of pandemic from newly emerging virulent strains. Current influenza vaccines *i.e.* trivalent inactivated vaccine (TIV) and live-attenuated vaccine (LAV) provide moderate protection which is greatly reduced or absent in some seasons (Osterholm *et al*., 2012). Pandemics of flu in the past have indicated that these vaccines were not efficient against new virulent strains. The Spanish flu was influenza pandemic which is considered as the most serious pandemics of recorded history (Table 3.1). Lists of flu pandemics are given in following Table 3.1.

Table 3.1: Details of known influenza pandemics in past

Name of Pandemics	Year	No. of Deaths	Subtype involved
Asiatic or Russian Flu	1889-1890	1 million	Possibly H3N8 or H2N2
Spanish Flu	1918-1920	20-100 million	H1N1
Asian Flu	1957-1958	1-1.5 million	H2N2
Hong Kong Flu	1968-1969	0.75-1 million	H3N2
Russian Flu	1977-1978	No accurate count	H1N1
2009 Flu pandemic	2009-2010	18000	H1N1/09

Current influenza vaccines generally consist of three influenza virus strains that the Centers for Disease Control and Prevention (CDC) determines to be the most predominant circulating strains for that flu season. Usually, all available influenza vaccines contain three strains of influenza virus which typically comprise two influenza A virus strains (H1N1 and H3N2) and one influenza B strains (Clayville, 2011).

To monitor the predominant circulating influenza strains and migration of influenza drift mutants for selection of appropriate influenza virus strains intended for vaccine, the Global Influenza Surveillance Network (GISN) was established by the World Health Organization (WHO) in 1947 (Cox *et al*., 1994). Now GISN comprises of 136 National Influenza Centres (NICs) at 106 countries, 5 WHO Collaborating Centres, 11 H5 Reference Laboratories and 4 Essential Regulatory Laboratories. GISN continues to expand its network to increase its global surveillance area and broaden its reach. Five WHO Collaborating Centres (Atlanta,

Beijing, London, Melbourne and Tokyo) receive collected samples (influenza specimens and isolates) from WHO NICs and Reference Laboratories for further study of sample for antigenic properties, genetic and drug sensitivity properties.

Influenza is one of the most studied viruses due to its involvement in occasional deadly pandemics and seasonal epidemics from ancient time. Around sixty thousand strains of influenza virus are reported worldwide and the number of strains continues to grow, millions of proteins sequences associated with strains and other information about location, antigenic properties and virulence demands appropriate storage and analysis system. Burgeoning amount of data generated through next generation sequencing technology and other high throughput experiments further demand the requirement of high end computational resources. Taking into account the importance of information resources for influenza virus, several computational resources were constructed which are not only having databases but also they provide the facility to do analysis on available data (*e.g.* Influenza Research database (IRD) and Influenza Virus Database (IVDB) (Chang *et al*., 2007)). Influenza web resources contain information about strains, nucleotides sequences, proteins sequences, epitopes sequences, location of strains, animal surveillance, sequence feature variant types, immune epitope data, 3D protein structures, host factor Data and other associated information (Chang *et al*., 2007; Squires *et al*., 2012). Details of the influenza-focused web-accessible resource are provided in Table 3.2.

Table 3.2: Web resources focused on influenza

Resource Name and Year of Establishment	**Web Link**	**Description**
Influenza Sequence & Epitope Database (ISED)	http://influenza.korea.ac.kr/ISED2/index_3.jsp	ISED catalogues the influenza sequence and epitope information obtained in countries worldwide and currently hosts a total of 50403 influenza A and 5215 influenza B virus sequence data including pandemic A/H1N1 2009 virus sequences collected from 42 countries. A total of 545 amantadine-resistant influenza virus sequences collected in Korea. ISED provides users with application tools to analyze

		sequence alignment and difference patterns, and allows users to visualize epitope matching structures (Yang *et al*., 2009).
Influenza Virus Database Systems (IVDS)	http://virusdb.czc.hokudai.ac.jp/	The IVDS presents influenza virus data obtained at the Graduate School of Veterinary Medicine and Research Center for Zoonosis Control, Hokkaido University, Japan, a member of the OIE Reference Laboratory for highly pathogenic avian influenza and low pathogenic avian influenza, for further global collaboration and data sharing.
Influenza Research Database (IRD)	http://www.fludb.org	This resource will contain avian and non-human mammalian influenza surveillance data, human clinical data associated with virus extracts, phenotypic characteristics of viruses isolated from extracts and all genomic and proteomic data available in public repositories for influenza viruses. The IRD provides a suite of tools for analysis of all types of influenza data and a personal work bench on which each user can store lists of important data selected from IRD resource (Squires *et al*., 2012).
Influenza Virus Resource (IVR)	http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html	The IVR obtains data from the NIAID Influenza Genome Sequencing Project as well as from GenBank. It combines sequences with tools for flu sequence analysis, annotation and finally submits the sequences to GenBank. In addition, it provides links to other resources that contain flu sequences, publications and general information about flu viruses (Bao *et al*., 2008).

Global Initiative on Sharing Avian Influenza Data (GISAID) EpiFlu Database	http://www.gisaid.org	This platform is designed and maintained for scientists from various disciplines in influenza research, including veterinary and human virology, bioinformatics, epidemiology, immunology and clinical analysis etc. From this resource one can find a series of services which pledge to offer the most complete information on influenza.
The Influenza Virus Database (IVDB)	http://influenza.psych.ac.cn/	The IVDB is an integrated information resource and analysis platform for genetic, genomic, and phylogenetic studies of influenza virus (Chang *et al*., 2007).
The OpenFlu database (OpenFluDB)	http://openflu.vital-it.ch/browse.php	It is a collaborative effort to share observations on the evolution of Influenza virus in both animals and humans. It contains genomic and protein sequences as well as epidemiological data from more than 25,000 isolates (Liechti *et al*., 2010).

Current influenza (TIV and LAV) vaccines provide protection by producing neutralizing antibodies against surface structural glycoproteins: hemagglutinin (HA) and neuraminidase (NA). But frequent mutations in these surface proteins results in escape of many virulent strains from antibody mediated immunity provided by vaccine strains (Fiore *et al*., 2009). Flu pandemic in 2009, and development of resistance strains to ribavirin and oseltamvir drugs which are first line of defense against influenza (Regoes and Bonhoeffer, 2006) reignited the hunt for UIV which can effectively counter the epidemics and pandemics caused by influenza virus.

Several research studies are exploring protein based vaccines for influenza. The main focus for the protein-based subunit vaccine design has been given to surface proteins such as hemagglutinin (HA) and neuraminidase (NA) but variable nature of these surface proteins led to the use of other influenza proteins for vaccines development. Influenza proteins known to provide immune responses include HA, NA, matrix protein, nucleoprotein and PB1 protein (Chen *et al*., 2011; Doucet *et al*., 2011; El Bakkouri *et al*., 2011; Košík *et al*., 2012; Sylte and Suarez, 2009). High sequence variations in this virus proteins lead to search for conserved

part of proteins which also contains immunogenic region (epitopes) for UIV design (Tan *et al*., 2011). Epitopes are the only region of protein required for immune response therefore, epitope has least possibility of side effects. In recent years, epitope-based vaccines have shown their effectiveness against human immune-deficiency virus (HIV), hepatitis B and influenza viruses in clinical studies (Atsmon *et al*., 2012; Engler *et al*., 2001; Gahery *et al*., 2006). Current approaches have been focusing on B-cell epitope (BCE) and/or T-cell epitope (TCE) mediated immune responses to develop the UIV (Goodman *et al*., 2011; Kaur *et al*., 2011). Single conserved ectodomain (epitope) of M2 (M2e) protein was developed by some company as UIV against influenza virus but mutations in middle part of the ectodomain eliminated its potential as UIV (Wang *et al*., 2009). In M2 protein the amino acid threonine and glutamic acid in positions five and six, respectively were responsible for the formation of escape variants from antibody response (Wang *et al*., 2009). Vaccine produced from single conserved TCE or BCE may not provide broad-specific protection but a vaccine developed from cocktail of a few conserved epitopes can take it closer to UIV (Atsmon *et al*., 2012). Multimeric-001 vaccine was developed by BiondVax with an objective to protect humans from seasonal as well as pandemic influenza strains which are developed due to genetic drifts and shifts. Multimeric-001, contains trimeric combination of 9 linear epitopes from three different proteins (Hemagglutinin (HA), nucleoprotein (NP) and matrix protein 1 (M1)) of influenza virus, has known to provide broad protection and is currently in phase 2 clinical trials (Atsmon *et al*., 2012). Recent positive outcomes of epitopes-based vaccines have opened an opportunity that the available enormous information (epitope, protein, and nucleotide sequences, immunogenic data, strains, etc.) on known epitopes (BCEs and TCEs) in public databases may be explored for the selection of epitopes and/or their combinations to design potential UIV. Therefore, a web resource named as EpiCombFlu has been developed which consists of "Epitope Information Resource" and "Epitopes Combination Explorer". The former is a database containing epitopes strains coverage and their immunogenic data while the latter explores combinations of epitopes for maximum strains coverage using forward selection algorithm (FSA).

3.2 Methods

3.2.1 *Data collection of proteins and epitopes of influenza virus*

All available sequences of 12 proteins (complete set of proteins encoded by influenza) HA, NA, NP, M1, M2, PB1, PB2, PB1-F2, NS1, NS2, PA and PA-X of influenza A virus were retrieved from influenza research database (http://www.fludb.org/). Redundant protein sequences within a strain were discarded while constricting protein sequences files for above mentioned 12 proteins. Different sequences of each individual protein (i.e. HA, NA, etc.) and their associated information such as strain name, subtype, country and host were stored. TCE and BCE "full data" files were downloaded from immune epitope database (IEDB) (http://www.immuneepitope.org/) and all epitopes source organism (influenza), host organism (human) and 'qualitative measurement' related to immune response (positive, positive-low, positive-intermediate or positive-high) from 11 influenza proteins were extracted separately. Individual epitopes along with their literature reference, IEDB Id, and information on types of immunogenic response and HLA allele (in case of TCE only) were stored in MySQL database as backend data for development of web resource. To avoid similar epitopes from same region of the same protein, the epitopes were clustered so that only one epitope will be selected from each cluster. CD-Hit (http://weizhong-lab.ucsd.edu/cd-hit/) was used for clustering of similar epitopes with threshold identity cut-off value 80%. Total number of non-overlapping clusters obtained in HA, NA, NP, M1, M2, PB1, PB2, PB1-F2, NS1, NS2 and PA proteins were 238, 72, 74, 33, 9, 48, 16, 4, 24, 10 and 20, respectively. Clustering of epitopes helps in covering different regions of a protein and provides option to avoid similar type of epitopes.

3.2.2 *Calculation of epitopes strain coverage among different strains and population coverage*

All epitopes were matched against respective proteins to determine their strain coverage. Exact sequence match is taken as conservation cut-off because there is immense inter-residues interaction in T-cell epitope so even a single amino acid substitution can alter interaction of other residues to TCR or HLA molecules (Rimmelzwaan *et al*., 2004). Data of all epitopes in terms of coverage in strains, vaccine strains, continent-wise strains, subtypes, and their host organism were calculated by using in-house programs. Individual strain coverage (ISC) of each epitope was computed as the number of strains containing the epitope

in their respective protein sequences and cumulative strain coverage (CSC) for combination of epitopes was determined as the number of all strains containing any epitope in their respective protein sequences. The strain coverage percentage of each epitope was calculated as ISC number multiplied by 100 and then divided by total number of strains where that protein sequence was available. Similarly coverage percentage of combination of epitopes was calculated as CSC number multiplied by 100 and then divided by total number of strains where any protein sequence was known. This coverage information was stored in MySQL database.

Population coverage of TCE based vaccine is crucial due to polymorphism of MHC molecules which display distinct peptide-binding specificity. TCE that binds to several HLA or common HLA supertype (most frequent HLA supertype in population) provides maximum population coverage. HLA supertypes of all TCEs (CTL and Th) were also stored. Global representative of the most frequent supertype of HLA-A and HLA-B (A1, A2, A3, B24, B7) were categorized to provide population coverage information (Sidney *et al*., 2008).

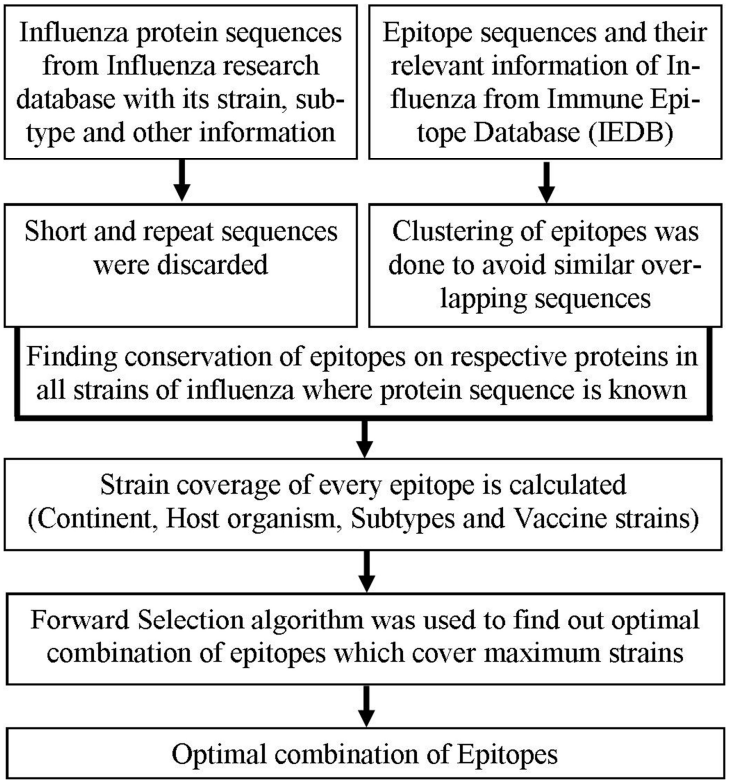


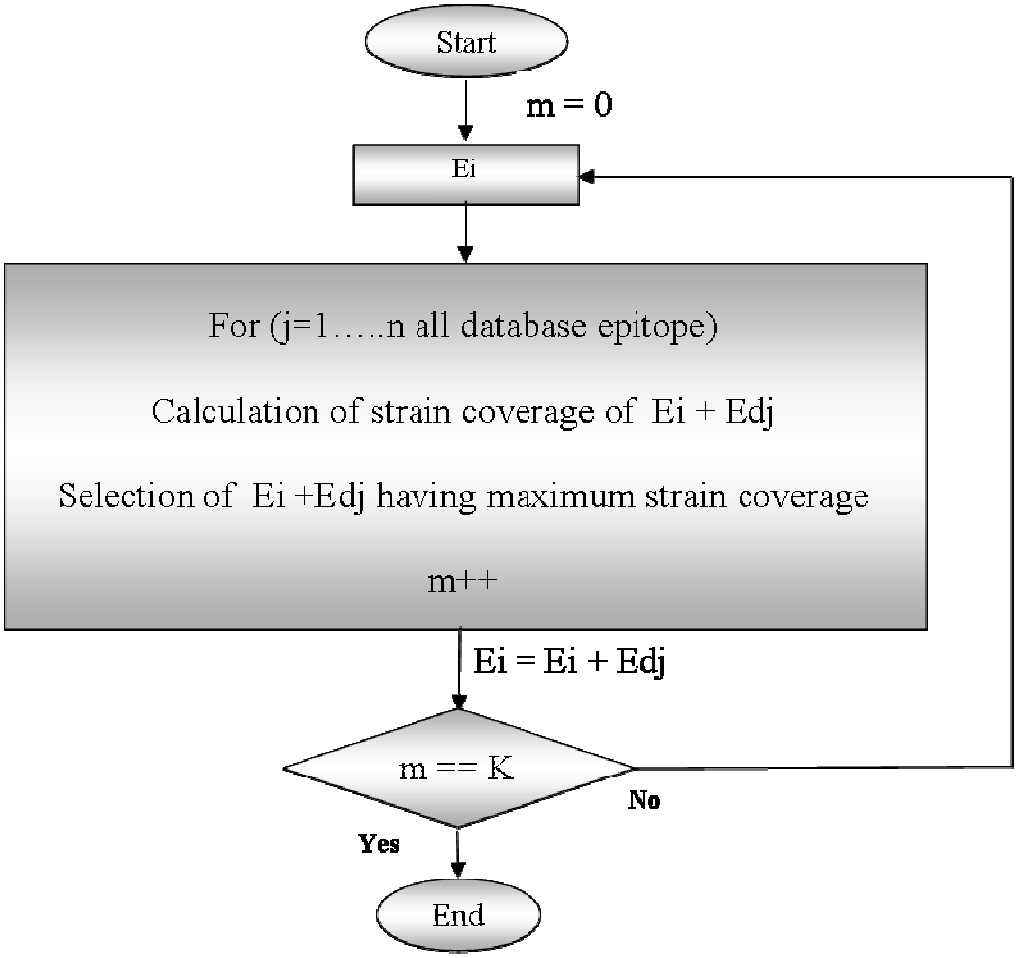
Figure 3.2: Methodology followed in EpiCombFlu

3.2.3 ***Database: Epitope Information Resource (EIR)***

Immunogenic information of epitopes and their coverage data were stored in MySQL database as backend data. The user interface (web pages) was designed in hypertext markup language (HTML), and PERL and PHP languages were used in developing server side program, which retrieve information from MySQL database for processing user's request. EpiCombFlu's resource can be searched by protein's name, continent's name, host organism's name and/or epitope's type. The results can be refined and tabulated (downloadable in .xls format).

3.2.4 ***Forward selection algorithm (FSA) for finding optimal combination of epitopes (Epitope Combination Explorer)***

The main component of EpiCombFlu is "Epitope Combination Explorer", which implements FSA to explore different combinations of epitopes for UIV. Evaluation of all combinations of epitopes covering approximately all global strains of influenza can result in combinatorial explosion. Therefore, FSA was developed and implemented in the web server to determine optimal combination of epitopes covering almost all global strains. By default, FSA takes epitope with maximum strain coverage as initial one. To select the next one (epitope), it determines cumulative strains coverage (CSC) for combination of new and initial epitopes, and the epitope (irrespective of protein) with maximum strain coverage is included into the combination. Similar method is followed to include third epitope into the combination, and the algorithm will incorporate epitopes in an iterative manner till user mentioned numbers of epitopes are added in the combination. The FSA can also take initial epitope(s) from user and then adds new epitopes (according to FSA) to provide optimal combination for maximum strains coverage. The stepwise execution of FSA and its output is provided in Figure 3.3 and 'Tutorial' (http://14.139.240.55/vaccine/tutorial.html).



Ei- Initial epitope or initially selected epitopes

Ed- All known epitopes of influenza stored in our database

K- User provided number of epitope included through FSA

Figure 3.3: Flow diagram of forward selection algorithm (FSA)

3.3 Results

Web resource EpiCombFlu provides compiled information on epitope's strain coverage, epitope-type, host-type, literature references, etc. so that user can evaluate the prospective epitope(s) for developing UIV. The web resource also provides analysis facility which adds epitope automatically on the basis of maximum strain coverage. The inclusion of information on strain and population coverage, and both humoral and cellular immune response features in EpiCombFlu are crucial for epitope-based broad spectrum vaccine (Ben-Yedidia and Arnon, 2005).

3.3.1 ***Description***

The web resource provides a user-friendly interface "Epitope Information Resource" to search and retrieve information related to influenza epitopes. Output of the search is presented in a tabular format which contains strain coverage and immunogenic information of epitopes. "Epitope Combination Explorer" evaluates different combinations of influenza epitopes using FSA. There are two options to submit epitope sequences into this module: 1. Selection of an epitope from dropdown menu of corresponding proteins (mouse on epitope sequence in the 'list box' displays a hover showing information about their immune responses, and coverage in strains (global, continents-wise and vaccine, strains), subtypes, and their host organism) and 2. Typing epitope sequence in provided text box Figure 3.4. The "Epitope Combination Explorer" provides similar information as "Epitope Information Resource" about conservation and immune response but for combination of selected epitopes.

3.3.2 ***Conservation of epitopes according to strain, sub-type and host-type***

Epitope from HA, GLFGAIAGFI, has maximum strain coverage and is conserved in 27402 strains. Top 10 epitopes from HA protein has conservation range in between 27402 to 13761 strains. Some epitopes were highly conserved and more than seventy epitopes were conserved in more than 10000 strains and twelve epitopes were found to be conserved in more than 20000 strains (Table 3.3). In web resource, strain coverage information of each epitope is hovered so that user can select conserved epitopes. Strain coverage data of top 10 epitopes of each protein is given in Table 3.3.

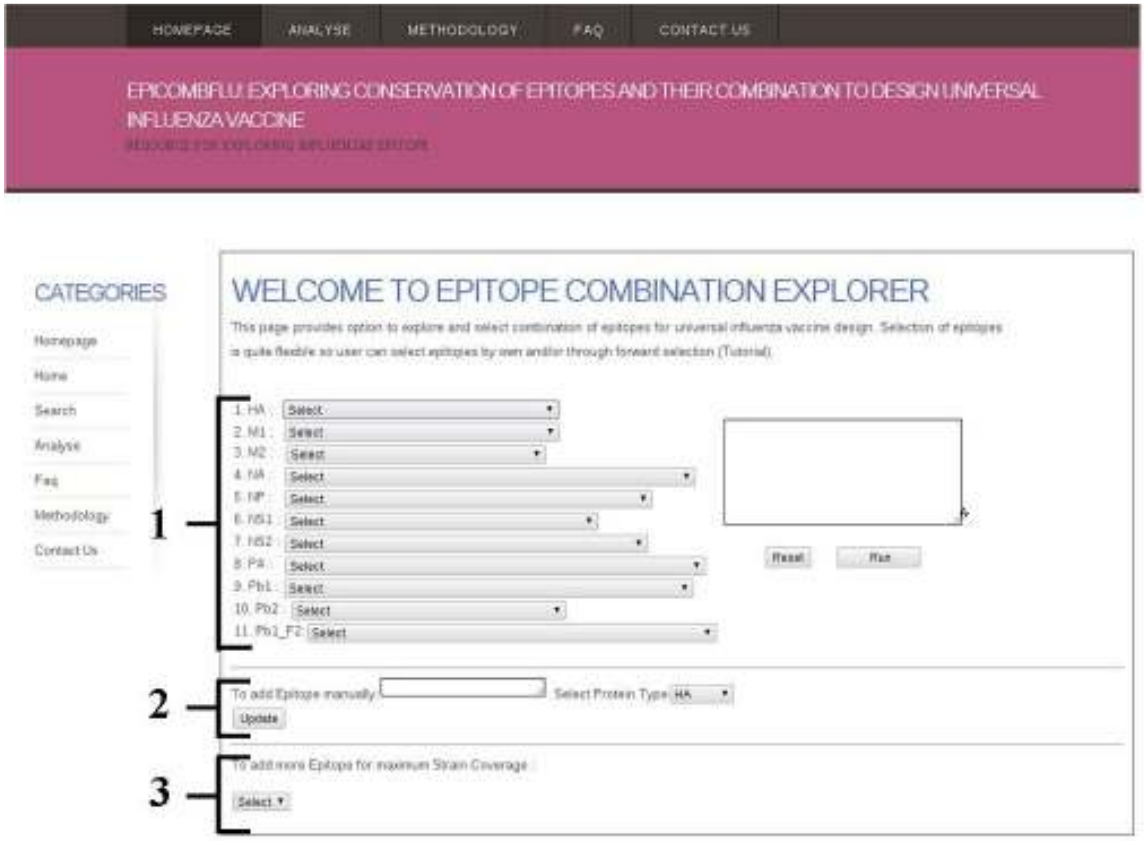


Figure 3.4: Job submission page of EpiCombFlu depicting three different methods for submission of a job. i) Selection of an epitope from drop down menu of corresponding protein, ii) Typing epitope sequence in provided text box and iii) To add epitopes through FSA

Table 3.3: Top 10 epitopes according to strain coverage data of each eleven influenza proteins

Protein Name	Epitope Id	Epitope Type	All Strain	Asia	North America	Europe	Africa	South America	Oceania	Human	Avian	Swine	MHC Allele Name
HA	20837	T-cell	27402	11382	9655	3966	1198	594	568	12846	11610	2400	HLA-A*02:01|HLA-A*02:02|HLA-A*02:03|HLA-A*02:06|HLA-A*68:02
HA	20838	B-cell	27303	11319	9641	3947	1197	592	568	12792	11571	2394	Not Applicable
HA	62335	T-cell	17276	5699	6835	3287	338	641	459	14067	21	3172	H-2-IEd
HA	11335	T-cell	17048	5805	6505	3289	335	638	460	14150	22	2861	HLA-DRB5*01:01
HA	151026	B-cell	16979	5777	6476	3277	335	638	460	14112	22	2831	Not Applicable
HA	95880	T-cell	16562	5325	7082	2746	337	582	475	13031	254	3249	HLA-Class II allele undetermined
HA	151039	B-cell	16556	5331	7085	2730	337	583	475	13023	254	3249	Not Applicable
HA	95458	T-cell	15846	4998	6832	2674	283	575	469	12555	79	3199	HLA-Class II allele undetermined
HA	150978	B-cell	13807	5045	4970	2541	264	519	455	11411	223	2145	Not Applicable
HA	79809	T-cell	13761	4512	5655	2587	104	478	415	10482	1442	1781	HLA-DRA*01:01/DRB1*01:01|/DRB1*04:01
M1	18170	T-cell	24550	9034	10978	2564	479	253	1211	12794	8272	2675	HLA-DR4
M1	79905	T-cell	24528	9021	10971	2562	479	253	1211	12794	8264	2661	HLA-DRA*01:01/DRB1*04:01
M1	33844	T-cell	24518	8947	11016	2584	470	253	1216	12790	8234	2691	HLA-A3
M1	4349	T-cell	24507	9105	10878	2537	481	260	1209	12846	8326	2675	HLA-B35|HLA-B*35:01

M1	97192	T-cell	24495	9097	10876	2535	481	260	1209	12845	8321	2669	HLA-A1
M1	144210	T-cell	24453	9077	10872	2528	481	259	1199	12825	8300	2669	HLA-A*03:01|HLA-A*11:01|HLA-A*31:01|HLA-A*33:01|HLA-A*68:01
M1	97255	T-cell	24097	8655	10939	2542	467	252	1211	12714	7951	2653	HLA-Class I allele undetermined
M1	129032	T-cell	24033	8936	10450	2688	473	255	1198	12937	7958	2598	HLA-DRB5
M1	97730	T-cell	24015	8822	10756	2490	475	245	1195	12503	8243	2625	HLA-Class II allele undetermined
M1	65112	T-cell	23996	8814	10755	2488	472	244	1191	12497	8229	2629	HLA-DRB1*04:04
M2	144461	T-cell	14594	4939	6822	2022	301	201	293	5381	6752	2149	HLA-A*03:01|HLA-A*11:01|HLA-A*31:01|HLA-A*33:01|HLA-A*68:01
M2	97562	T-cell	14143	4358	7131	1194	346	64	1024	6294	6153	1218	—
M2	97518	T-cell	6573	2821	2457	506	116	40	620	6540	4	29	—
M2	68383	T-cell	6134	1629	2899	563	123	42	866	5809	77	248	HLA-B44
M2	97545	T-cell	6166	2440	3080	488	22	20	104	127	5706	149	HLA-Class I allele undetermined
M2	97727	T-cell	5461	1396	2542	523	119	40	833	5389	10	62	—
M2	59318	B-cell	4929	1295	2416	431	117	40	622	4876	4	49	Not Applicable
M2	97749	B-cell	4847	1765	2525	435	18	12	81	103	4526	64	Not Applicable
M2	97544	T-cell	4085	1161	1708	398	105	7	703	4054	1	30	HLA-Class I allele undetermined
M2	128914	T-cell	2100	858	935	80	12	26	185	2098	2	0	HLA-DRB1*04:04

NA	127810	T-cell	14624	5883	4770	2705	568	273	409	9764	3549	1189	HLA-DRB1*01:01|HLA-DRB1*07:01|HLA-DRB1*11:01
NA	135949	T-cell	14049	5601	4540	2665	562	273	394	9718	3051	1164	HLA-DRB1*01:01|HLA-DRB1*07:01|HLA-DRB5*01:01
NA	135932	T-cell	13111	5700	4157	2436	141	273	395	9813	2417	775	HLA-DRB1*11:01|HLA-DRB5*01:01
NA	62486	T-cell	12088	4928	4083	2045	386	233	397	8645	2632	711	HLA-A24
NA	135909	T-cell	11710	4918	3525	2268	548	251	193	7227	2955	1441	HLA-DRB1*15:01
NA	136008	T-cell	11681	4916	3486	2270	552	251	193	7320	2993	1274	HLA-DRB1*15:01
NA	130337	T-cell	11553	4644	4017	1903	363	233	377	8380	2425	652	—
NA	135946	T-cell	10808	4675	2953	2275	455	250	194	7286	2975	461	HLA-DRB1*04:01|HLA-DRB1*15:01
NA	135938	T-cell	10801	5010	2591	2302	442	249	194	7360	2666	685	HLA-DRB1*04:01
NA	135937	T-cell	10722	4588	2933	2294	458	250	193	7251	2918	469	HLA-DRB1*04:01|HLA-DRB1*15:01
NP	9745	T-cell	18914	6457	8631	2041	321	230	1206	9634	7389	1393	HLA-Class II allele undetermined
NP	21255	T-cell	18906	6452	8640	2043	321	227	1195	9620	7399	1403	HLA-A*02:01
NP	38688	T-cell	18883	6463	8614	2029	321	229	1199	9609	7392	1384	HLA-B*07:02|HLA-B*54:01
NP	97772	T-cell	18873	6444	8617	2033	327	230	1194	9632	7348	1395	HLA-B44|HLA-B*18:01
NP	27126	T-cell	18867	6419	8613	2043	331	227	1206	9634	7335	1397	HLA-B8
NP	49220	T-cell	18843	6441	8605	2020	320	229	1200	9605	7374	1368	HLA-DRB1*01:01

NP	27283	T-cell	18800	6390	8577	2153	217	227	1204	9838	7058	1404	HLA-A3|HLA-A28
NP	67439	T-cell	18626	6461	8350	2053	318	224	1192	9611	7269	1389	HLA-Class II allele undetermined
NP	56698	T-cell	18648	6010	8620	2222	335	229	1202	9852	6867	1427	HLA-DQ5
NP	164408	T-cell	18586	6441	8335	2048	318	224	1192	9606	7243	1381	HLA-Class I allele undetermined
NS1	10014	T-cell	14918	5953	6083	1845	329	213	467	6518	6416	1427	HLA-DR3
NS1	19312	T-cell	13187	5225	5295	1711	315	189	428	5485	6015	1137	HLA-B44
NS1	130242	T-cell	10648	3849	4459	990	340	48	942	5105	5259	108	HLA-DRB1*01:01
NS1	144293	T-cell	9136	4190	3433	880	250	53	306	2022	6208	377	HLA-A*02:01|HLA-A*02:02|HLA-A*02:03|HLA-A*02:06|HLA-A*68:02
NS1	128294	T-cell	7202	2011	3504	1062	71	197	350	5876	226	991	HLA-DRB1*03:01
NS1	97170	B-cell	6802	3402	2416	554	273	19	119	494	5713	68	Not Applicable
NS1	128504	T-cell	6195	2641	2687	535	35	27	257	1425	4423	182	HLA-DRB1*11:01
NS1	27467	T-cell	4761	1226	2808	410	33	36	236	1372	2933	32	HLA-Class I allele undetermined
NS1	2014	T-cell	4758	1226	2806	409	33	36	236	1372	2932	32	HLA-A*02:01|HLA-A2
NS1	128445	T-cell	4724	1212	2788	408	32	36	236	1371	2902	32	HLA-A*02:01
NS2	97405	T-cell	17089	6037	7385	1976	325	193	1148	9627	5643	1354	HLA-A2
NS2	97195	T-cell	13738	5088	6237	1665	277	163	290	4723	7110	1309	HLA-A2

NS2	144440	T-cell	10280	3161	5037	880	232	58	891	4862	4269	747	HLA-B*18:01|HLA-B*40:01|HLA-B*40:02|HLA-B*44:03|HLA-B*45:01|HLA-B*44:02
NS2	148935	T-cell	6564	2633	3048	500	34	51	277	1808	4239	57	HLA-DR|
NS2	41785	T-cell	6442	2537	2858	655	67	41	267	1832	4261	212	HLA-Class II allele undetermined
NS2	97498	T-cell	3742	1102	1527	368	64	5	674	3301	399	40	HLA-Class II allele undetermined
NS2	97757	T-cell	3641	853	1576	443	60	10	691	3575	16	50	—
NS2	129425	T-cell	1340	312	751	50	1	28	198	1286	52	1	—
NS2	129280	T-cell	1297	313	709	53	1	29	190	1250	36	10	—
NS2	128860	T-cell	1239	295	690	43	1	29	181	1206	29	2	—
PA	17119	T-cell	18194	5514	8989	1955	311	220	1178	9214	7125	1325	HLA-A*02:01|HLA-A*02:02|HLA-A*02:03|HLA-A*02:06|HLA-A*68:02|HLA-A2
PA	129181	T-cell	18150	5870	8722	1824	317	212	1177	8931	7484	1245	—
PA	76533	T-cell	18076	5458	8949	1949	296	222	1175	9204	7042	1305	HLA-A*24:02|HLA-A*23:01
PA	148589	T-cell	18040	5733	8758	1836	296	216	1173	9121	7198	1238	HLA-DR
PA	97503	T-cell	17899	5735	8622	1823	309	219	1164	9026	7278	1217	HLA-Class I allele undetermined
PA	128477	T-cell	17873	5689	8626	1867	291	218	1163	9067	7302	1185	—
PA	144377	T-cell	17562	5121	8785	1929	309	222	1170	9158	6885	999	HLA-A*03:01|HLA-A*11:01|HLA-A*31:01|HLA-A*68:01

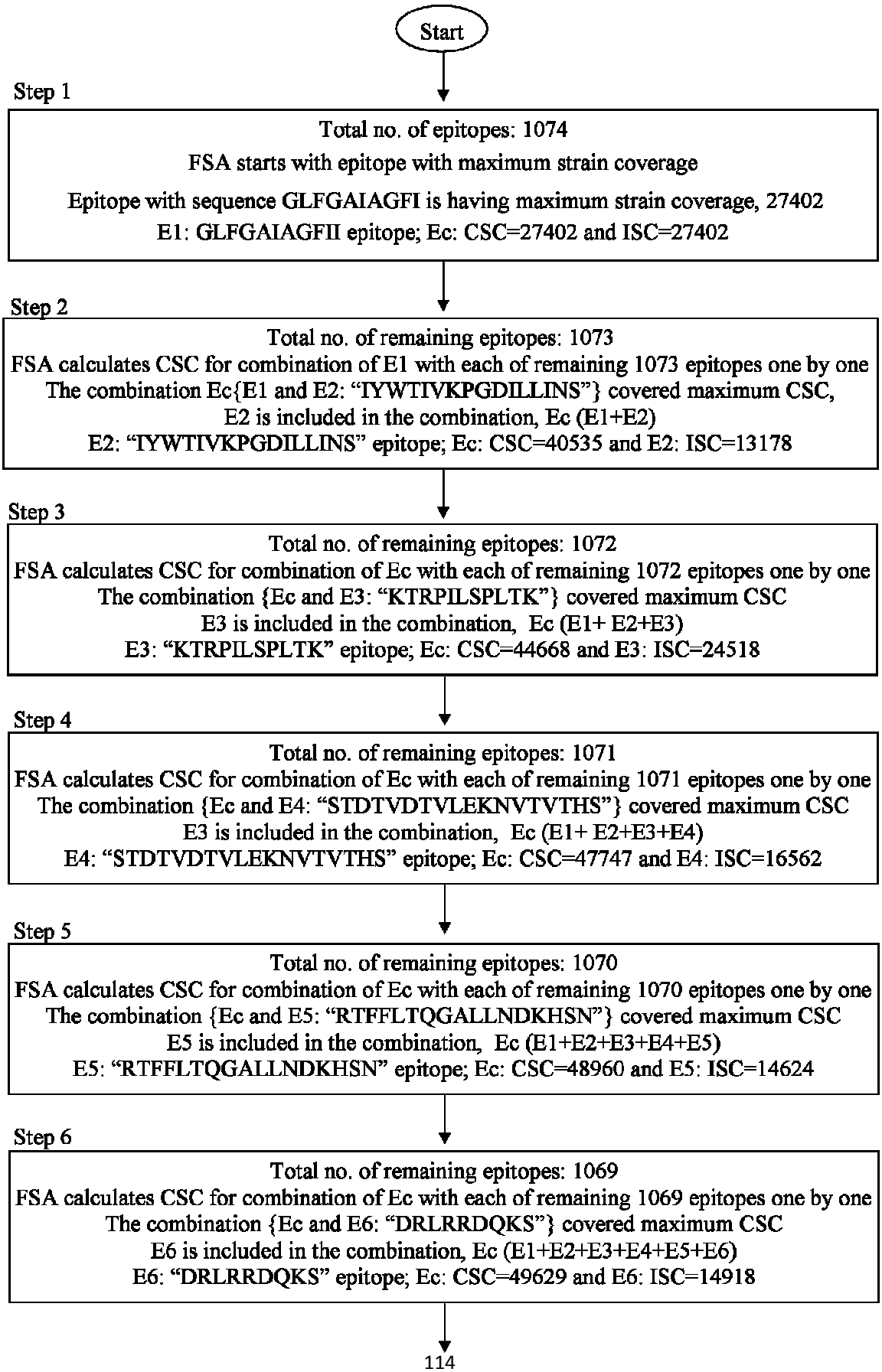
PA	97623	T-cell	17290	5680	8256	1895	238	188	1007	8013	7445	1305	HLA-DR
PA	62180	T-cell	16607	5255	8405	1718	311	216	676	8110	6976	1046	HLA-B8
PA	148968	T-cell	14735	5278	6953	1473	274	214	521	6988	6454	1029	HLA-DR
PB1	75755	T-cell	18702	6168	8941	1860	320	220	1166	9201	7656	1331	HLA-A*01:01|HLA-A*26:01|HLA-A*30:02|HLA-A*29:02
PB1	10514	T-cell	18698	6219	8883	1860	322	221	1166	9280	7592	1318	HLA-A26
PB1	97682	T-cell	18596	6094	8922	1850	317	220	1166	9144	7620	1321	HLA-Class I allele undetermined
PB1	165648	T-cell	18603	6084	8948	1832	327	222	1163	9073	7644	1360	HLA-B27
PB1	127207	T-cell	18565	6158	8842	1839	318	219	1162	9248	7520	1291	HLA-DRB1*04:01
PB1	45001	T-cell	18573	6112	8864	1834	341	221	1171	9101	7633	1344	HLA-A*02:01|HLA-A*02:02|HLA-A*02:03|HLA-A*02:06|HLA-A*68:02
PB1	6174	T-cell	18558	6064	8926	1826	330	222	1163	9061	7634	1347	HLA-B44
PB1	97309	T-cell	18554	6108	8862	1833	329	221	1171	9099	7617	1343	HLA-A2
PB1	42143	T-cell	18543	6111	8852	1831	327	221	1171	9102	7613	1343	HLA-A2
PB1	97655	T-cell	18514	6090	8846	1829	327	221	1171	9094	7595	1342	HLA-DR
PB2	97519	T-cell	1113	97	486	142	3	0	385	1103	0	10	—
PB2	144475	T-cell	178	172	0	0	0	0	0	15	151	0	HLA-A*03:01|HLA-A*11:01|HLA-A*31:01|HLA-A*33:01|HLA-A*68:01

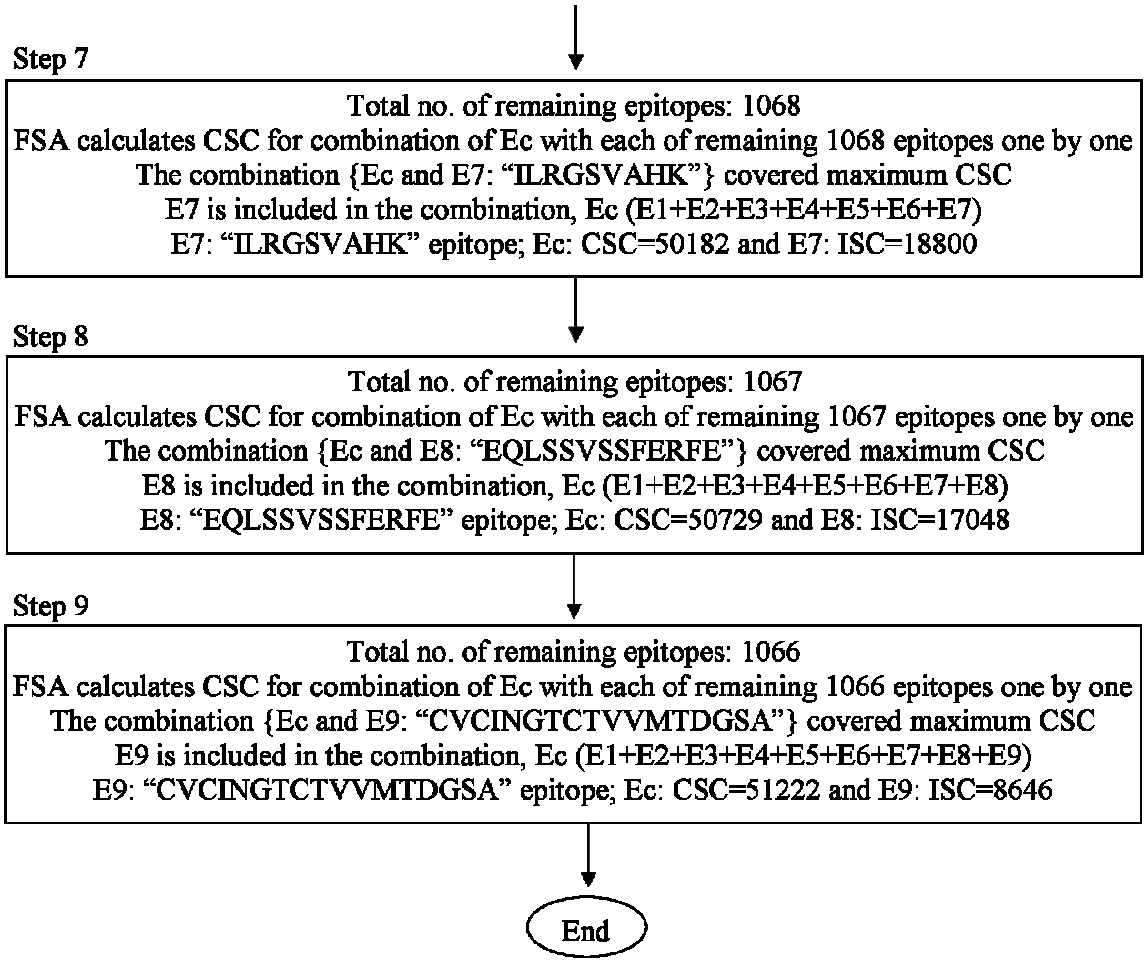
PB2	130187	T-cell	2	2	0	0	0	0	0	2	0	0	—
PB2	97779	T-cell	2	2	0	0	0	0	0	2	0	0	HLA-Class I allele undetermined
PB2	164387	T-cell	18293	5870	8802	1911	298	220	1168	9224	7303	1293	HLA-B27
PB2	148683	T-cell	18070	5575	8808	1972	309	217	1168	9278	7005	1312	HLA-DR
PB2	144526	T-cell	17827	5404	8777	1922	311	218	1168	9200	6844	1302	HLA-A*03:01|HLA-A*11:01|HLA-A*31:01|HLA-A*68:01
PB2	97446	T-cell	17411	5200	8615	1882	313	215	1159	9031	6684	1248	HLA-DR
PB2	97696	T-cell	17033	5128	8408	1815	296	213	1147	9041	6591	952	—
PB2	173541	T-cell	16417	5375	7848	1719	290	177	981	8055	7084	828	—
PB1f2	129123	T-cell	2780	630	1273	141	40	5	689	2748	0	32	—
PB1f2	97574	B-cell	2700	920	1376	67	49	5	279	1742	808	19	Not Applicable
PB1f2	97462	B-cell	182	182	0	0	0	0	0	24	153	0	Not Applicable

3.3.3 ***Performance of the FSA for UIV design***

When the epitope with maximum strain coverage, GLFGAIAGFI, was used as an initial epitope for FSA and all other epitopes were added automatically based on maximum cumulative strains coverage of combined epitopes (Refer Section 3.2.4, Figure. 3.5), nine epitopes were selected from five proteins (4 HA, 2 NA, 1 M1, 1 NP and 1 NS1) covering 51222 strains out of total 57414 strains of influenza A virus (Table 3.4). Epitopes 1, 3 and 7 provide CTL immune responses. Epitopes bind to MHC II alleles were taken as Th-cell epitopes. Interestingly, other six epitopes are known to induce Th immune response which is crucial to encounter viral infection (Tan *et al*., 2011). First and seventh CTL and second and fifth Th epitopes are known to bind with multiple HLA supertypes indicating their applicability to global human population.

Epitope-based vaccine, Multimeric-001, is composed of nine epitopes (four BCE and one Th epitope from HA protein, two CTL and one Th from NP protein, and one peptide that contains both BCE and CTL epitope from M1 protein) and epitopes forming this vaccine have coverage of only 30848 strains of influenza A virus (Table 3.4) in comparison to 51222 strains coverage of 9 epitopes combination discovered in the present work by FSA. Besides coverage, FSA discovered combination of epitopes is expected to induce better immune response and population coverage which is important in vaccine design against influenza virus.


Start
Step 1
Total no. of epitopes: 1074
FSA starts with epitope with maximum strain coverage
Epitope with sequence GLFGAIAGFI is having maximum strain coverage, 27402
E1: GLFGAIAGFII epitope; Ec: CSC=27402 and ISC=27402
Step 2
Total no. of remaining epitopes: 1073
FSA calculates CSC for combination of E1 with each of remaining 1073 epitopes one by one
The combination Ec{E1 and E2: "IYWTIVKPGDILLINS"} covered maximum CSC,
E2 is included in the combination, Ec (E1+E2)
E2: "IYWTIVKPGDILLINS" epitope; Ec: CSC=40535 and E2: ISC=13178
Step 3
Total no. of remaining epitopes: 1072
FSA calculates CSC for combination of Ec with each of remaining 1072 epitopes one by one
The combination {Ec and E3: "KTRPILSPLTK"} covered maximum CSC
E3 is included in the combination, Ec (E1+ E2+E3)
E3: "KTRPILSPLTK" epitope; Ec: CSC=44668 and E3: ISC=24518
Step 4
Total no. of remaining epitopes: 1071
FSA calculates CSC for combination of Ec with each of remaining 1071 epitopes one by one
The combination {Ec and E4: "STDTVDTVLEKNVTVTHS"} covered maximum CSC
E3 is included in the combination, Ec (E1+ E2+E3+E4)
E4: "STDTVDTVLEKNVTVTHS" epitope; Ec: CSC=47747 and E4: ISC=16562
Step 5
Total no. of remaining epitopes: 1070
FSA calculates CSC for combination of Ec with each of remaining 1070 epitopes one by one
The combination {Ec and E5: "RTFFLTQGALLNDKHSN"} covered maximum CSC
E5 is included in the combination, Ec (E1+E2+E3+E4+E5)
E5: "RTFFLTQGALLNDKHSN" epitope; Ec: CSC=48960 and E5: ISC=14624
Step 6
Total no. of remaining epitopes: 1069
FSA calculates CSC for combination of Ec with each of remaining 1069 epitopes one by one
The combination {Ec and E6: "DRLRRDQKS"} covered maximum CSC
E6 is included in the combination, Ec (E1+E2+E3+E4+E5+E6)
E6: "DRLRRDQKS" epitope; Ec: CSC=49629 and E6: ISC=14918



- End result of the FSA including sequences of epitopes, Epitope ID, CSC, ISC, MHC-allele, etc. has been provided in Table 3.4 (Case-II).
- The above flow diagram is also effective for the cases of user selected epitope or set of epitopes. The user provided epitope(s) are incorporated in the starting steps of flow diagram, and the moment user wishes FSA to combine epitopes, the FSA computes CSC for the combination of already selected epitopes, Ec with each of remaining epitopes one by one. The combination which provides maximum CSC is included into the combination. The detailed procedure related execution of FSA has been provided in the tutorials. Few outputs using user selected epitope(s) has been provided in the Table 3.4 (details in Table 3.5-3.12).

Figure 3.5: Stepwise execution of FSA started through epitope (GLFGAIAGFI) having maximum strain coverage (refer section 3.3.3)

3.3.4 ***Output***

Output results of "Epitope Combination Explorer" (ECE) were designed to display in tabular format for easy interpretation. First row of result table shows information about the selected combination of epitopes and all other rows represent the similar information but for individual epitopes taken into combination. Output table is having four columns. First column represents the epitope ID (IEDB ID); second column contains the description of epitope or combination of epitopes in terms of conservation (according global strains, continent-wise strain, vaccine strains, host-wise, and subtypes), epitope type (T-cell, B-cell or Both), MHC allele name (in case of T-cell) and cluster ID of epitope; and third and fourth columns contain pie chart according to continent wise strain coverage of epitope or combination of epitopes (fourth column) and global strain coverage of epitope or combination of epitopes (fifth column). Sample output for combination of two epitopes is provided in Figure 3.6.

Result of Selected Epitopes and their Combination for Universal Influenza Vaccine

Download result in text

Colour scheme: ASIA[RED] NORTH AMERICA[GREEN] EUROPE[DARK BLUE] AFRICA[YELLOW] SOUTH AMERICA[PINK] OCEANIA[BLUE]

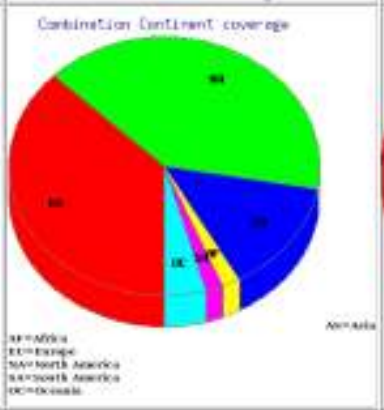
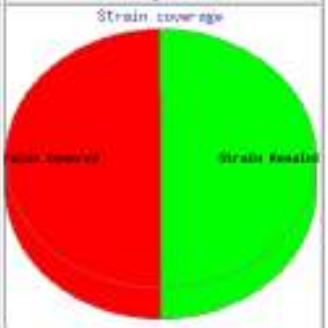
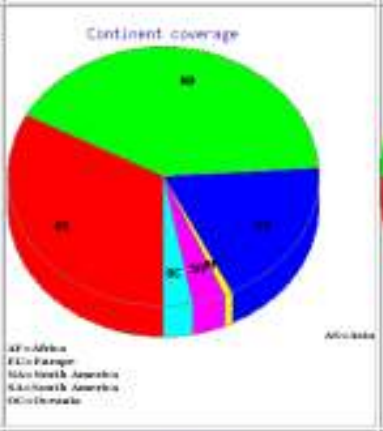
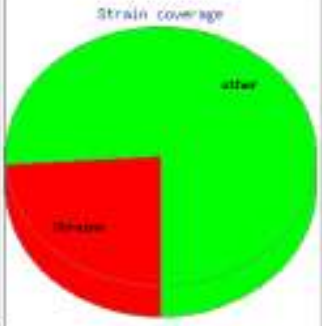
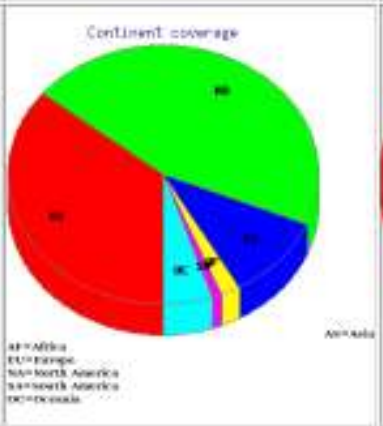
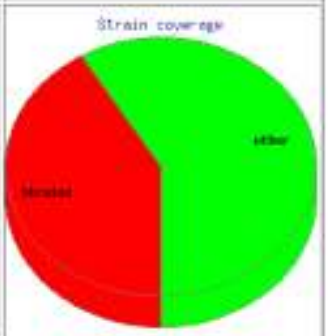
Epitope Combination or ID	Description	Continent Wise Strain Coverage	All Strain Coverage
Combination_Information	COMBINATION Strains: 28639 49.88% NORTH_AMERICA: 11548 40.37% ASIA: 10697 37.40% EUROPE: 4055 14.18% OCEANIA: 1244 4.35% AFRICA: 544 1.90% SOUTH_AMERICA: 515 1.80% VACCINE: 29 HUMAN: 16854 60.47% AVIAN: 8113 29.11% SWINE: 2906 10.43% H1N1: 13138 H3N2: 5678 H5N1: 2141		
EPITOPE_IEDB_ID:1510	EPITOPE_IEDB_ID:151075 151075 Epitope_Sequence:YNAELLVLLENE Cluster_ID:Cluster_ID:HA46 Epitope_Type:B-Cell MHC_Allele_Name:Not_Applicable Strains: 13752 23.95% NORTH_AMERICA: 5642 41.06% ASIA: 4517 32.87% EUROPE: 2585 18.81% SOUTH_AMERICA: 480 3.49% OCEANIA: 415 3.02% AFRICA: 103 0.75% VACCINE: 10 HUMAN: 10489 76.57% SWINE: 1771 12.93% AVIAN: 1438 10.50% H1N1: 11654 H1N2: 723 H6N2: 414 H6N1: 327		
EPITOPE_IEDB_ID:1579	EPITOPE_IEDB_ID:1579 Epitope_Sequence:AGKNTDLEA Cluster_ID:Cluster_ID:M11 Epitope_Type:T-Cell MHC_Allele_Name:HLA-DRB1*07:01 Strains: 23742 41.35% NORTH_AMERICA: 10788 45.50% ASIA: 8491 35.81% EUROPE: 2520 10.63% OCEANIA: 1200 5.06% AFRICA: 463 1.95% SOUTH_AMERICA: 250 1.05% VACCINE: 28 HUMAN: 12610 54.80% AVIAN: 7787 33.84% SWINE: 2615 11.36% H1N1: 8732 H3N2: 5678 H5N1: 2141 H3N8: 843		

Figure 3.6: Output result of "Epitope Combination Explorer" calculated from the combination of two epitopes

Table 3.4: Comparative analysis of strains coverage and their continent-wise distribution for different combination of epitopes (Multimeric-001 and EpiCombFlu using FSA)*

Case	Methods of epitopes combination	All Strains Coverage	Asia	Africa	Europe	North America	Oceania	South America	Human	Avian	Swine
I.	Multimeric-001	30848	17213	925	5931	17095	2805	1142	21485	6983	1418
II.	FSA selection	51222	21242	1704	8061	17145	1802	1205	31184	13670	5167
III.	20 length epitope from HA as initial epitope and subsequent 8 epitopes were selected automatically through FSA	51057	21204	1706	8043	17034	1804	1203	31130	13579	5143
IV.	20 length epitope from HA as initial epitope and subsequently only more than 10 lengths epitopes were selected automatically through FSA	50916	21043	1662	7983	16939	1787	1439	31430	13312	5092
V.	First BCE was selected manually and subsequent 8 epitopes were selected automatically by FSA	51178	21201	1703	8059	17145	1802	1205	31165	13646	5166
VI.	First two BCEs were selected manually and subsequent 7 epitopes were selected automatically by FSA	51141	21203	1691	8048	17129	1802	1205	31167	13610	5160
VII.	First four BCEs (3 epitopes from HA and 1 form M1) were selected manually and subsequent 5 epitopes were selected automatically by FSA	51122	21199	1691	8039	17123	1802	1205	31154	13610	5155
VIII.	First four BCEs (low strain	48999	20418	1643	7447	16494	1755	1182	29918	13164	4858

	coverage) were selected manually and subsequent 5 epitopes were selected automatically by FSA										
IX.	Only Th epitopes were selected through FSA	49713	20860	1687	7615	16547	1749	1194	30697	12815	5101

* The Multimeric-001 was developed by BiondVax (Atsmon *et al*., 2012) and EpiCombFlu method has been developed in the present work.

Table 3.5: Case II: FSA identified nine epitopes, and their information related to epitope ID, CSC, ISC, different host-strain coverage, immune response and MHC allele distribution is provided.

Epitope number	Epitope sequences	[$]Epitope ID	[$]CSC	[$]ISC	Human strains	Avian strain	Swine strain	Type of Immune response	MHC Allele
E1	GLFGAIAGFI	20837	27402	27402	12846	11610	2400	T-cell	HLA-A*02:01 HLA-A*02:02 HLA-A*02:03 HLA-A*02:06 HLA-A*68:02
E2	IYWTIVKPGDILLINS	29727	40535	13178	12730	3	411	Th-cell	HLA-DRB1*07:01 HLA-DRB1*15:01
E3	KTRPILSPLTK	33844	44668	24518	12790	8234	2691	T-cell	HLA-A3
E4	STDTVDTVLEKNVTVTHS	95880	47747	16562	13031	254	3249	Th-cell	HLA-Class II allele undetermined
E5	RTFFLTQGALLNDKHSN	127810	48960	14624	9764	3549	1189	Th-cell	HLA-DRB1*01:01 HLA-DRB1*07:01 HLA-DRB1*11:01
E6	DRLRRDQKS	10014	49629	14918	6518	6416	1427	Th-cell	HLA-DR3
E7	ILRGSVAHK	27283	50182	18800	9838	7058	1404	T-cell	HLA-A3 HLA-A28
E8	EQLSSVSSFERFE	113375	50729	17048	14150	22	2861	Th-cell	HLA-DRB5*01:01
E9	CVCINGTCTVVMTDGSA	127615	51222	8646	5625	1790	1203	Th-cell	DRB1*01:01

[$] Refer section 3.3.3 and Table 3.4. Epitope ID related to IEDB. CSC: cumulative strain coverage of combined epitopes and ISC: individual strain coverage.

Table 3.6: Case III: FSA 20 length epitope from HA was selected as initial epitope and subsequent 8 epitopes were selected automatically through FSA. Nine epitopes information related to epitope ID, CSC, ISC, different host-strain coverage, immune response and MHC allele distribution is provided.

Epitope number	Epitope sequences	[$]Epitope ID	[$]CSC	[$]ISC	Human strains	Avian strain	Swine strain	Type of Immune response	MHC Allele
E1	CYPYDVPDYASLRSLVAS SG	138761	12682	12682	11526	945	129	Th-cell	HLA-DRB5*01:01 HLA-DRB1*04:01
E2	GLFGAIAGFI	20837	39020	27402	12846	11610	2400	T-cell	HLA-A*02:01 HLA-A*02:02 HLA-A*02:03 HLA-A*02:06 HLA-A*68:02
E3	KTRPILSPLTK	33844	43806	24518	12790	8234	2691	T-cell	HLA-A3
E4	STDTVDTVLEKNVTVTHS	95880	46885	16562	13031	254	3249	Th-cell	HLA-Class II allele undetermined
E5	RTFFLTQGALLNDKHSN	127810	48097	14624	9764	3549	1189	Th-cell	HLA-DRB1*01:01 HLA-DRB1*07:01 HLA-DRB1*11:01
E6	PGDILLINSTGNLIAPR	129567	49298	13065	12619	2	407	Th-cell	HLA-Class IIallele undetermined
E7	DRLRRDQKS	10014	49964	14918	6518	6416	1427	Th-cell	HLA-DR3
E8	EQLSSVSSFERFE	113375	50514	17048	14150	22	2861	Th-cell	HLA-DRB5*01:01
E9	ILRGSVAHK	27283	51057	18800	9838	7058	1404	T-cell	HLA-A3 HLA-A28

[$] Refer Table 3.4.. Epitope ID related to IEDB. CSC: cumulative strain coverage of combined epitopes and ISC: individual strain coverage

Table 3.7: Case IV: FSA 20 length epitope from HA was selected as initial epitope and subsequent 8 epitopes with 10 or more than 10 lengths were selected automatically through FSA. Nine epitopes information related to epitope ID, CSC, ISC, different host-strain coverage, immune response and MHC allele distribution is provided.

Epitope number	Epitope sequences	$Epitope ID	$CSC	$ISC	Human strains	Avian strain	Swine strain	Type of Immune response	MHC Allele
E1	CYPYDVPDYASLRSLVAS SG	138761	12682	12682	11526	945	129	Th-cell	HLA-DRB5*01:01 HLA-DRB1*04:01
E2	GLFGAIAGFI	20837	39020	27402	12846	11610	2400	T-cell	HLA-A*02:01 HLA-A*02:02 HLA-A*02:03 HLA-A*02:06 HLA-A*68:02
E3	KTRPILSPLTK	33844	43806	24518	12790	8234	2691	T-cell	HLA-A3
E4	STDTVDTVLEKNVTVTHS	95880	46885	16562	13031	254	3249	Th-cell	HLA-Class IIallele undetermined
E5	RTFFLTQGALLNDKHSN	127810	48097	14624	9764	3549	1189	Th-cell	HLA-DRB1*01:01 HLA-DRB1*07:01 HLA-DRB1*11:01
E6	PGDILLINSTGNLIAPR	129567	49298	13065	12619	2	407	Th-cell	HLA-Class IIallele undetermined
E7	EQLSSVSSFERFE	113375	49886	17048	14150	22	2861	Th-cell	HLA-DRB5*01:01
E8	LILRGSVAHKSCLPACVY	97448	50437	16963	8331	6868	1265	Th-cell	HLA-Class IIallele undetermined
E9	NQNLEYQIGYICSGIFG	135967	50916	7872	7182		521	Th-cell	DRB1*01:01

$ Refer Table 3.4. Epitope ID related to IEDB. CSC: cumulative strain coverage of combined epitopes and ISC: individual strain coverage

Table 3.8: Case V: FSA First BCE was selected manually and subsequent 8 epitopes were selected automatically by FSA. Nine epitopes information related to epitope ID, CSC, ISC, different host-strain coverage, immune response and MHC allele distribution is provided

Epitope number	Epitope sequences	[$]Epitope ID	[$]CSC	[$]ISC	Human strains	Avian strain	Swine strain	Type of Immune response	MHC Allele
E1	GLFGAIAGFIE	20838	27303	27303	12792	11571	2394	B-Cell	Not Applicable
E2	IYWTIVKPGDILLINS	29727	40436	13178	12730	3	411	Th-cell	HLA-DRB1*07:01 HLA-DRB1*15:01
E3	KTRPILSPLTK	33844	44584	24518	12790	8234	2691	T-cell	HLA-A3
E4	STDTVDTVLEKNVTVTHS	95880	47683	16562	13031	254	3249	Th-cell	HLA-Class IIallele undetermined
E5	RTFFLTQGALLNDKHSN	127810	48897	14624	9764	3549	1189	Th-cell	HLA-DRB1*01:01 HLA-DRB1*07:01 HLA-DRB1*11:01
E6	DRLRRDQKS	10014	49577	14918	6518	6416	1427	Th-cell	HLA-DR3
E7	EQLSSVSSFERFE	113375	50134	17048	14150	22	2861	Th-cell	HLA-DRB5*01:01
E8	ILRGSVAHK	27283	50684	18800	9838	7058	1404	T-cell	HLA-A3 HLA-A28
E9	CVCINGTCTVVMTDGSA	127615	51178	8646	5625	1790	1203	Th-cell	DRB1*01:01

[$] Refer Table 3.4. Epitope ID related to IEDB. CSC: cumulative strain coverage of combined epitopes and ISC: individual strain coverage

Table 3.9: Case VI: FSA First two BCEs were selected manually and subsequent 7 epitopes were selected automatically by FSA. Nine epitopes information related to epitope ID, CSC, ISC, different host-strain coverage, immune response and MHC allele distribution is provided.

Epitope number	Epitope sequences	[$]Epitope ID	[$]CSC	[$]ISC	Human strains	Avian strain	Swine strain	Type of Immune response	MHC Allele
E1	GLFGAIAGFIE	20838	27303	27303	12792	11571	2394	B-Cell	Not Applicable
E2	GILGFVFTL	20354	35790	23620	12785	7367	2664	T-cell/B-cell	HLA-A*02:01 HLA-A2 HLA-A*02:02 HLA-A*02:03 HLA-A*02:06 HLA-A*68:02 HLA-B*35:01
E3	IYWTIVKPGDILLINS	29727	44507	13178	12730	3	411	Th-cell	HLA-DRB1*07:01 HLA-DRB1*15:01
E4	STDTVDTVLEKNVTVTHS	95880	47606	16562	13031	254	3249	Th-cell	HLA-Class IIallele undetermined
E5	RTFFLTQGALLNDKHSN	127810	48830	14624	9764	3549	1189	Th-cell	HLA-DRB1*01:01 HLA-DRB1*07:01 HLA-DRB1*11:01
E6	DRLRRDQKS	10014	49530	14918	6518	6416	1427	Th-cell	HLA-DR3
E7	EQLSSVSSFERFE	113375	50087	17048	14150	22	2861	Th-cell	HLA-DRB5*01:01
E8	ILRGSVAHK	27283	50640	18800	9838	7058	1404	T-cell	HLA-A3 HLA-A28
E9	CVCINGTCTVVMTDGSA	127615	51141	8646	5625	1790	1203	Th-cell	DRB1*01:01

[$] Refer Table 3.4. Epitope ID related to IEDB. CSC: cumulative strain coverage of combined epitopes and ISC: individual strain coverage

Table 3.10: Case VII: FSA first four BCEs (3 epitopes from HA and 1 form M1) were selected manually and subsequent 7 epitopes were selected automatically by FSA. Nine epitopes information related to epitope ID, CSC, ISC, different host-strain coverage, immune response and MHC allele distribution is provided.

Epitope number	Epitope sequences	$Epitope ID	$CSC	$ISC	Human strains	Avian strain	Swine strain	Type of Immune response	MHC Allele
E1	GLFGAIAGFIE	20838	27303	27303	12792	11571	2394	B-Cell	Not Applicable
E2	LREQLSSVSSFERFE	151026	35790	16979	14112	22	2831	B-Cell	Not Applicable
E3	GILGFVFTL	20354	38858	23620	12785	7367	2664	T-cell/B-Cell	HLA-A*02:01 HLA-A2 HLA-A*02:02 HLA-A*02:03 HLA-A*02:06 HLA-A*68:02 HLA-B*35:01
E4	NSTDTVDTVLEKNVT	151039	39515	16556	13023	254	3249	B-Cell	Not Applicable
E5	IYWTIVKPGDILLINS	29727	48211	13178	12730	3	411	Th-cell	HLA-DRB1*07:01 HLA-DRB1*15:01
E6	RTFFLTQGALLNDKHSN	127810	49416	14624	9764	3549	1189	Th-cell	HLA-DRB1*01:01 HLA-DRB1*07:01 HLA-DRB1*11:01
E7	DRLRRDQKS	10014	50068	14918	6518	6416	1427	Th-cell	HLA-DR3
E8	ILRGSVAHK	27283	50621	18800	9838	7058	1404	T-cell	HLA-A3 HLA-A28
E9	CVCINGTCTVVMTDGSA	127615	51122	8646	5625	1790	1203	Th-cell	DRB1*01:01

$ Refer Table 3.4. Epitope ID related to IEDB. CSC: cumulative strain coverage of combined epitopes and ISC: individual strain coverage

Table 3.11: Case VIII: FSA first four BCEs (with low strain coverage) were selected manually and subsequent 5 epitopes were selected automatically by FSA. Nine epitopes information related to epitope ID, CSC, ISC, different host-strain coverage, immune response and MHC allele distribution is provided.

Epitope number	Epitope sequences	$Epitope ID	$CSC	$ISC	Human strains	Avian strain	Swine strain	Type of Immune response	MHC Allele
E1	HKIFKMEKGKVVKSVELD APNYHY	97383	453	453	56	380	0	B-Cell	Not Applicable
E2	RVTVSTRRSQQTIIPNIG	56434	1542	1089	78	934	7	B-Cell	Not Applicable
E3	MSLLTEVETYVLSIIPSGPL KAEIAQKLEDVFAGKNTD LEALMEWLKTRPI	97507	2894	1629	118	1446	12	B-Cell	Not Applicable
E4	RKKRGLFGAIAGFIE	163243	4920	3185	320	2752	24	B-Cell	Not Applicable
E5	GLFGAIAGFI	20837	27601	27402	12846	11610	2400	T-cell	HLA-A*02:01 HLA-A*02:02 HLA-A*02:03 HLA-A*02:06 HLA-A*68:02
E6	IYWTIVKPGDILLINS	29727	40733	13178	12730	3	411	Th-cell	HLA-DRB1*07:01 HLA-DRB1*15:01
E7	KTRPILSPLTK	33844	44721	24518	12790	8234	2691	T-cell	HLA-A3
E8	STDTVDTVLEKNVTVTHS	95880	47800	16562	13031	254	3249	Th-cell	HLA-Class IIallele undetermined

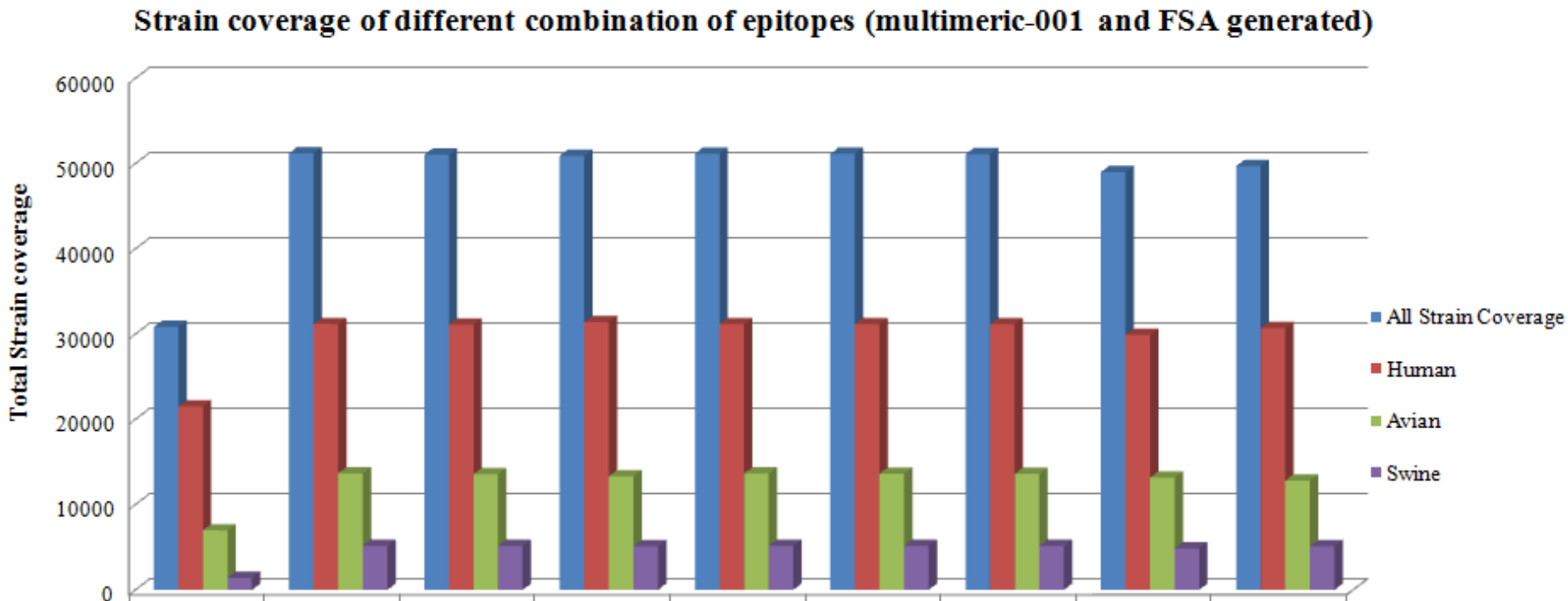
E9	RTFFLTQGALLNDKHSN	127810	48999	14624	9764	3549	1189	Th-cell	HLA-DRB1*01:01 HLA-DRB1*07:01 HLA-DRB1*11:01

[$] Refer Table 3.4. Epitope ID related to IEDB. CSC: cumulative strain coverage of combined epitopes and ISC: individual strain coverage

Table 3.12: Case IX: FSA only Th epitopes were selected through FSA. Nine epitopes information related to epitope ID, CSC, ISC, different host-strain coverage, immune response and MHC allele distribution is provided.

Epitope number	Epitope sequences	[$]Epitope ID	[$]CSC	[$]ISC	Human strains	Avian strain	Swine strain	Type of Immune response	MHC Allele
E1	FVFTLTVPSER	18170	24550	24550	12794	8272	2675	Th-Cell	HLA-DR4
E2	EQLSSVSSFERFE	113375	33632	17048	14150	22	2861	Th-cell	HLA-DRB5*01:01
E3	IYWTIVKPGDILLINS	29727	42331	13178	12730	3	411	Th-cell	HLA-DRB1*07:01 HLA-DRB1*15:01
E4	NIHPLTIGECPKYVK	139558	44672	4884	374	4365	28	Th-cell	HLA-Class IIallele undetermined
E5	RTFFLTQGALLNDKHSN	127810	46197	14624	9764	3549	1189	Th-cell	HLA-DRB1*01:01 HLA-DRB1*07:01 HLA-DRB1*11:01
E6	NAELLVLLENQKTLDEHDAN	152823	47577	2218	9	2115	32	Th-cell	HLA-DRB1*01:01 HLA-DRB1*03:01 HLA-DRB1*04:01
E7	STDTVDTVLEKNVTVTHS	95880	48356	16562	13031	254	3249	Th-cell	HLA-Class IIallele undetermined
E8	DRLRRDQKS	10014	49065	14918	6518	6416	1427	Th-cell	HLA-DR3
E9	GFAPFSKDNSIRLSAGG	127658	49713	9280	5666	2209	1303	Th-cell	HLA-DRB1*03:01 HLA-DRB1*04:01

[$] Refer Table 3.4. Epitope ID related to IEDB. CSC: cumulative strain coverage of combined epitopes and ISC: individual strain coverage



	Multimeric-001 (Case I)	FSA (Case II)	FSA (Case III)	FSA (Case IV)	FSA (Case V)	FSA (Case VI)	FSA (Case VII)	FSA (Case VIII)	FSA (Case IX)
All Strain Coverage	30848	51222	51057	50916	51178	51141	51122	48999	49713
Human	21485	31184	31130	31430	31165	31167	31154	29918	30697
Avian	6983	13670	13579	13312	13646	13610	13610	13164	12815
Swine	1418	5167	5143	5092	5166	5160	5155	4858	5101

Figure 3.7: Comparison of combination of epitopes used in mulmeric-001 and generated through FSA in different conditions (Table 3.4)

3.4 DISCUSSION

The EpiCombFlu resource has been developed to assist vaccinologists for producing epitope-based UIVs. Epitope Information Resource (EIR) implemented in EpiCombFlu having information about known epitopes which are providing immune response in humans. EIR incorporated calculated conservation of all epitopes according to location (continent-wise), vaccine strain, subtypes and host. In light of high variability of influenza virus these information are essential for universal vaccine design. EIR also incorporate other immunogenic information of epitopes which is important in epitopes based vaccine design. Furthermore some of the accumulated epitopes were highly conserved (more than seventy epitopes were conserved in more than 10000 strains and twelve epitopes were conserved in more than 20000 strains). Thorough conserved epitopes with their other associated information important for vaccine design (such as immune response, MHC allele, protein name, literature reference, epitope sequence, host, location, and conservations) EIR expected to bridge the gap of antigenic variability in influenza virus which is the major hurdle in influenza vaccine design. None of known epitopes of influenza was conserved in all known influenza virus strains therefore only through combination of conserved epitopes the strain coverage of influenza virus can maximize. Recent research of Multimeric-001 also emphasizes the importance of combination of conserved epitopes in UIV design. Therefore, the most important feature of EpiCombFlu is the "Epitope Combination Explorer (ECE)" design to explore vaccine potential of different combination of epitopes. ECE provides facility to design different combination of epitopes manually and/or automatically. Automatically combination of epitopes is generated through adding up the epitope(s) according to FSA. Different combinations of epitopes (potential UIVs with optimum strain coverage) can be identified with the help of FSA with different initial epitope(s). The flexible options to design combination of epitopes is implemented in ECE which provide unique facility to user so, that user can use his/her domain knowledge and power of computational algorithm simultaneously in selection different combination of epitopes for UIV. The combination of epitopes can be expressed as synthetic protein and its UIV potential could be checked. Since evolutionary conserved epitopes are from functionally important parts of proteins, therefore, these immunogenic parts are expected to be retained by new pandemic strains (McMurry *et al*., 2008). The developed vaccine containing these epitopes is anticipated to prevent or mitigate infection by pandemic strains.

The performance of EpiCombFlu server does not depend on initially selected epitopes. Many independent studies were also carried out to verify whether selection of different

combination of epitopes can alter the strain coverage or not (Table 3.4). In first case, instead of taking epitope with maximum strain coverage as initial epitope, a 20 length epitope from HA was taken and FSA was used to find out combination of epitopes with optimal strain coverage. Similarly in second case, the same 20 length epitope was taken as initial epitope and then only 10 or more length epitopes were used subsequently for combination by FSA (Table 3.4). The FSA method was able to find combination of 9 epitopes having around 90% strains coverage. Even when 4 BCEs with low strain coverage were used as initial epitopes then also web resource (FSA) was able to get 85 percent strain coverage with addition of five epitopes. This outcome of FSA justifies strains coverage is not dependent on selection of initial epitopes. Even if the user may like to include one or more epitopes (i.e. known to be highly immunogenic or preferred epitopes of user) for developing the UIV then also the web resource (FSA) can easily combine other database epitopes to get maximum strain coverage (Table 3.4). When combination of epitopes used in multimeric-001 compared with FSA generated combinations, than all FSA generated combination were having extensively more strain coverage in all cases such as avian, human, swine and global strains (Figure 3.7). Comparison result again justified the importance of EpiCombFlu for universal influenza vaccine design. EpiCombFlu server is freely accessible for research and education purpose at http://14.139.240.55/influenza/home.html. Output results and associated data and search result of database "Epitope Information Resource" can be downloaded from available links. Currently, standalone version of tool is not available.

EpiCombFlu is developed on linear epitopes and further its efficiency may be improved by considering discontinuous epitopes. Conservation calculation of structural epitopes should not be like linear epitopes and three dimensional structures of epitopes must be taken in consideration. Although structure of majority of epitopes are not available but inclusion of structural epitopes may further improve the reliability of methodology followed in EpiCombFlu.

CONCLUSIONS

The prediction of broad-specific vaccine candidates and design of universal epitope-based vaccines are anticipated to overcome the limitations of current vaccines. In this work, known immunogenic and biological information, available sequence data and other associated information were used to develop novel methods which were implemented subsequently as cybertools to predict better vaccine candidates. The Jenner-Predict server has been developed to predict potential protein vaccine candidates (PVCs) and also to provide their vaccine potential with an objective of assisting subunit vaccine development. The web server was validated on independent and diverse datasets, where it outperformed other PVC prediction tools. Its performance substantiated that the proteins involved in host-pathogen interactions and pathogenesis are better criteria than methods based on machine learning or adhesin-likeliness. The method based on host-pathogen interaction predicts less number of PVCs in a proteome with high prediction accuracy which confirms its reliability. The vaccine potential of PVCs is evaluated in terms of their possible immunogenicity by comparing with known immunogenic epitopes, absence of autoimmunity and conservation in different strains. Mapping of known epitopes from immune epitope database (IEDB) on PVCs increases the probability of a protein to be immunogenic. Comparison of these PVCs with human proteome sequences reduces the chance of their failure due to autoimmunity. Conservation of PVCs in pathogenic strains provides crucial information on their broad-specificities. Since the web server provides prioritized PVCs, few prospective proteins from a proteome could be taken for experimental evaluation to identify subunit vaccine candidates.

Web resource, EpiCombFlu has been developed, which consists of the "Epitope Information Resource" (database containing epitopes' strain coverage and their immunogenic data) and the "Epitope Combination Explorer" to discover combinations of epitopes for maximum strains' coverage using forward selection algorithm (FSA). Comparative study has shown that it provides better combination of epitopes in comparison with all other reported and/or analyzed epitopes combinations for universal influenza vaccine (UIV). Combinations of nine epitopes through FSA were conserved in ~90% global influenza A virus strains justify the potential of EpiCombFlu in UIV design. Simulation studies have shown that the FSA did not depend on initial epitope selection and was effective in providing combination of epitopes covering optimum (more than 85%) global strain coverage. This resource is expected to accelerate the development of universal vaccine against influenza. The cybertools developed in the present work are not organism specific *i.e.* the methodology used in Jenner-predict can be extended to

protozoa, fungi, parasites, etc. and similarly, the principle employed in EpiCombFlu can be expanded to other highly variable pathogenic viruses in which epitopes and proteins sequences or several strains are known.

BIBLIOGRAPHY

Achouak W., Heulin T., "Multiple facets of bacterial porins," *FEMS microbiology letters,* vol. 199, pp. 1-7, 2001.

Aidinis V., Chandras C., Manoloukos M., Thanassopoulou A., Kranidioti K., Armaka M., Douni E., Kontoyiannis D., Zouberakis M., Kollias G., "MUGEN mouse database; animal models of human immunological diseases," *Nucleic acids research,* vol. 36, pp. D1048-D1054, 2008.

Aithal A., Sharma A., Joshi S., Raghava G.P., Varshney G.C., "PolysacDB: A Database of Microbial Polysaccharide Antigens and Their Antibodies," *PloS one,* vol. 7, p. e34613, 2012.

Akahata, W., Yang, Z. Y., Andersen, H., Sun, S., Holdaway, H. A., Kong, W. P., Lewis M.G., Higgs S., Rossmann M.G., Rao S., Nabel G.J., "A virus-like particle vaccine for epidemic Chikungunya virus protects nonhuman primates against infection," *Nature medicine*, vol. 16, p. 334-338, 2010.

Alexander J.E., Lock R.A., Peeters C., Poolman J.T., Andrew P.W., Mitchell T.J., Hansman D., Paton J.C., "Immunization of mice with pneumolysin toxoid confers a significant degree of protection against at least nine serotypes of *Streptococcus pneumoniae*," *Infection and immunity,* vol. 62, pp. 5683-5688, 1994.

Ansari H.R., Flower D.R., Raghava G., "AntigenDB: an immunoinformatics database of pathogen antigens," *Nucleic acids research,* vol. 38, pp. D847-D853, 2010.

Ansari H.R., Raghava G.P., "Identification of conformational B-cell Epitopes in an antigen from its primary sequence," *Immunome research,* vol. 6, p. 6, 2010.

Ariel N., Zvi A., Grosfeld H., Gat O., Inbar Y., Velan B., Cohen S., Shafferman A., "Search for potential vaccine candidate open reading frames in the *Bacillus anthracis* virulence plasmid pXO1: in silico and in vitro screening," *Infection and immunity,* vol. 70, pp. 6817-6827, 2002.

Atsmon J., Kate-Ilovitz E., Shaikevich D., Singer Y., Volokhov I., Haim K.Y., Ben-Yedidia T., "Safety and Immunogenicity of Multimeric-001—a Novel Universal Influenza Vaccine," *Journal of clinical immunology,* vol. 32, pp. 595-603, 2012.

Bao Y., Bolotov P., Dernovoy D., Kiryutin B., Zaslavsky L., Tatusova T., Ostell J., Lipman D., "The influenza virus resource at the National Center for Biotechnology Information," *Journal of virology,* vol. 82, pp. 596-601, 2008.

Barh D., Barve N., Gupta K., Chandra S., Jain N., Tiwari S., Leon-Sicairos N., Canizalez-Roman A., dos Santos A.R., Hassan S.S., "Exoproteome and Secretome Derived Broad Spectrum Novel Drug and Vaccine Candidates in *Vibrio cholerae* Targeted by Piper betel Derived Compounds," *PloS one,* vol. 8, p. e52773, 2013.

Barouch D.H., Liu J., Peter L., Abbink P., Iampietro M.J., Cheung A., Alter G., Chung A., Dugast A-S., Frahm N., "Characterization of humoral and cellular immune responses elicited by a recombinant adenovirus serotype 26 HIV-1 Env vaccine in healthy adults (IPCAVD 001)," *Journal of Infectious Diseases,* vol. 207, pp. 248-256, 2013.

Barrio A.M., Soeria-Atmadja D., Nistér A., Gustafsson M.G., Hammerling U., Bongcam-Rudloff E., "EVALLER: a web server for in silico assessment of potential protein allergenicity," *Nucleic acids research,* vol. 35, pp. W694-W700, 2007.

Ben-Yedidia T., Arnon R., "Towards an epitope-based human vaccine for influenza," *Human vaccines,* vol. 1, pp. 95-101, 2005.

Berry A.M., Lock R.A., Hansman D., Paton J.C., "Contribution of autolysin to virulence of *Streptococcus pneumoniae,*" *Infection and immunity,* vol. 57, pp. 2324-2330, 1989.

Bhasin M., Raghava G., "A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes," *Journal of biosciences,* vol. 32, pp. 31-42, 2007.

Bhasin M., Raghava G., "Prediction of CTL epitopes using QM, SVM and ANN techniques," *Vaccine,* vol. 22, pp. 3195-3204, 2004.

Bhasin M., Raghava G., "SVM based method for predicting HLA-DRB1* 0401 binding peptides in an antigen sequence," *Bioinformatics,* vol. 20, pp. 421-423, 2004.

Bhasin M., Singh H., Raghava G., "MHCBN: a comprehensive database of MHC binding and non-binding peptides," *Bioinformatics,* vol. 19, pp. 665-666, 2003.

Blythe M.J., Flower D.R., "Benchmarking B cell epitope prediction: underperformance of existing methods," *Protein Science,* vol. 14, pp. 246-248, 2005.

Bongaerts R.J., Heinz H-P., Hadding U., Zysk G., "Antigenicity, Expression, and Molecular Characterization of Surface-Located Pullulanase of *Streptococcus pneumoniae*," *Infection and immunity,* vol. 68, pp. 7141-7143, 2000.

Briles D.E., Ades E., Paton J.C., Sampson J.S., Carlone G.M., Huebner R.C., Virolainen A., Swiatlo E., Hollingshead S.K., "Intranasal immunization of mice with a mixture of the pneumococcal proteins PsaA and PspA is highly protective against nasopharyngeal carriage of *Streptococcus pneumoniae*," *Infection and immunity,* vol. 68, pp. 796-800, 2000.

Brown J.S., Ogunniyi A.D., Woodrow M.C., Holden D.W., Paton J.C., "Immunization with components of two iron uptake ABC transporters protects mice against systemic *Streptococcus pneumoniae* infection," *Infection and immunity,* vol. 69, pp. 6702-6706, 2001.

Brusic V., Rudy G., Harrison L.C., "MHCPEP, a database of MHC-binding peptides: update 1997," *Nucleic acids research,* vol. 26, pp. 368-371, 1998.

Caines M.E., Zhu H., Vuckovic M., Willis L.M., Withers S.G., Wakarchuk W.W., Strynadka N.C., "The Structural Basis for T-antigen Hydrolysis by *Streptococcus pneumoniae* A TARGET FOR STRUCTURE-BASED VACCINE DESIGN," *Journal of Biological Chemistry,* vol. 283, pp. 31279-31283, 2008.

Cao J., Gong Y., Cai B., Feng W., Wu Y., Li L., Zou Y., Ying B., Wang L., "Modulation of human bronchial epithelial cells by pneumococcal choline binding protein A," *Human Immunology,* vol. 72, pp. 37-46, 2011.

Chailyan A., Tramontano A., Marcatili P., "A database of immunoglobulins with integrated tools: DIGIT," *Nucleic acids research,* vol. 40, pp. D1230-D1234, 2012.

Chakravarti D.N., Fiske M.J., Fletcher L.D., Zagursky R.J., "Application of genomics and proteomics for identification of bacterial gene products as potential vaccine candidates," *Vaccine,* vol. 19, pp. 601-612, 2000.

Chang S., Zhang J., Liao X., Zhu X., Wang D., Zhu J., Feng T., Zhu B., Gao G.F., Wang J., "Influenza Virus Database (IVDB): an integrated information resource and analysis platform for influenza virus research," *Nucleic acids research,* vol. 35, pp. D376-D380, 2007.

Chapman R., Chege G., Shephard E., Stutz H., Williamson A-L., "Recombinant *Mycobacterium bovis* BCG as an HIV vaccine vector," *Current HIV research,* vol. 8, p. 282, 2010.

Chen J-R., Ma C., Wong C-H., "Vaccine design of hemagglutinin glycoprotein against influenza," *Trends in biotechnology,* vol. 29, pp. 426-434, 2011.

Chen Y-S., Hsiao Y-S., Lin H-H., Yen C-M., Chen S-C., Chen Y-L., "Immunogenicity and anti-*Burkholderia pseudomallei* activity in Balb/c mice immunized with plasmid DNA encoding flagellin," *Vaccine,* vol. 24, pp. 750-758, 2006.

Chen Z., Zhou M., Gao X., Zhang G., Ren G., Gnanadurai C.W., Fu Z.F., He B., "A novel rabies vaccine based on a recombinant parainfluenza virus 5 expressing rabies virus glycoprotein," *Journal of virology,* vol. 87, pp. 2986-2993, 2013.

Clayville L.R., "Influenza update: A review of currently available vaccines," *Pharmacy and Therapeutics,* vol. 36, p. 659, 2011.

Cohen T., Moise L., Martin W., De Groot A.S., "Immunoinformatics: The Next Step in Vaccine Design," in *Infectious Disease Informatics*, ed: Springer, 2010, pp. 223-244.

Cox N.J., Brammer T.L., Regnery H.L. Regnery, "Influenza: global surveillance for epidemic and pandemic variants," *European journal of epidemiology,* vol. 10, pp. 467-470, 1994.

Curtiss III R., Xin W., Li Y., Kong W., Wanda S-Y., Gunn B., Wang S., "New Technologies in Using Recombinant Attenuated *Salmonella* Vaccine Vectors," *Critical Reviews™ in Immunology,* vol. 30, 2010.

D'Agostino M., Martin G., "The bioscience revolution & the biological weapons threat: levers & interventions," *Globalization and health,* vol. 5, p. 3, 2009.

De Groot A.S., Moise L., "New tools, new approaches and new ideas for vaccine development," 2007.

Del Val M., Schlicht H-J., Volkmer H., Messerle M., Reddehase M.J., Koszinowski U.H., "Protection against lethal cytomegalovirus infection by a recombinant vaccine containing a single nonameric T-cell epitope," *Journal of virology,* vol. 65, pp. 3641-3646, 1991.

Denich K., Blyn L., Craiu A., Braaten B., Hardy J., Low D., O'Hanley P., "DNA sequences of three papA genes from uropathogenic *Escherichia coli* strains: evidence of structural and serological conservation," *Infection and immunity,* vol. 59, pp. 3849-3858, 1991.

Doan, L. X., Li, M., Chen, C., & Yao, Q., "Virus-like particles as HIV-1 vaccines," *Reviews in medical virology*, vol. 15, p. 75-88, 2005.

Dönnes P., Elofsson A., "Prediction of MHC class I binding peptides, using SVMHC," *BMC Bioinformatics,* vol. 3, p. 25, 2002.

Doucet J-D., Forget M-A., Grange C., Rouxel R.N., Arbour N., Von Messling V., Lapointe R., "Endogenously expressed matrix protein M1 and nucleoprotein of influenza A are efficiently presented by class I and class II major histocompatibility complexes," *Journal of general virology,* vol. 92, pp. 1162-1171, 2011.

Doytchinova I.A., Flower D.R., "VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines," *BMC Bioinformatics,* vol. 8, p. 4, 2007.

Durant L., Metais A., Soulama-Mouze C., Genevard J-M., Nassif X., Escaich S., "Identification of Candidates for a Subunit Vaccine against Extraintestinal Pathogenic *Escherichia coli*," *Infection and immunity,* vol. 75, pp. 1916-1925, April 1, 2007 2007.

Easton D.M., Smith A., Gallego S.G., Foxwell A.R., Cripps A.W., Kyd J.M., "Characterization of a novel porin protein from *Moraxella catarrhalis* and identification of an immunodominant surface loop," *Journal of bacteriology,* vol. 187, pp. 6528-6535, 2005.

Ehreth J., "The global value of vaccination," *Vaccine,* vol. 21, pp. 596-600, 2003.

Einhorn M., Anderson E., Weinberg G., Granoff P., Granoff D., "Immunogenicity in infants of *Haemophilus influenzae* type B polysaccharide in a conjugate vaccine with *Neisseria meningitidis* outer-membrane protein," *The Lancet,* vol. 328, pp. 299-302, 1986.

El Bakkouri K., Descamps F., De Filette M., Smet A., Festjens E., Birkett A., Van Rooijen N., Verbeek S., Fiers W., Saelens X., "Universal vaccine based on ectodomain of matrix protein 2 of influenza A: Fc receptors and alveolar macrophages mediate protection," *The Journal of Immunology,* vol. 186, pp. 1022-1031, 2011.

El-Adhami W., Kyd J.M., Bastin D.A., Cripps A.W., "Characterization of the gene encoding a 26-kilodalton protein (OMP26) from nontypeable *Haemophilus influenzae* and immune responses to the recombinant protein," *Infection and immunity,* vol. 67, pp. 1935-1942, 1999.

EL-Manzalawy Y., Dobbs D., Honavar V., "Predicting linear B-cell epitopes using string kernels," *Journal of Molecular Recognition,* vol. 21, pp. 243-255, 2008.

Enders J.F., Weller T.H., Robbins F.C., "Cultivation of the Lansing strain of poliomyelitis virus in cultures of various human embryonic tissues," *Science,* vol. 109, pp. 85-87, 1949.

Engler O.B., Dai W.J., Sette A., Hunziker I.P., Reichen J., Pichler W.J., Cerny A., "Peptide vaccines against hepatitis B virus: from animal model to human studies," *Molecular immunology,* vol. 38, pp. 457-465, 2001.

Fiers M.W., Kleter G.A., Nijland H., Peijnenburg A.A., Nap J.P., van Ham R.C., "Allermatch™, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines," *BMC Bioinformatics,* vol. 5, p. 133, 2004.

Fiore A., Bridges C., Cox N., "Seasonal Influenza Vaccines," in *Vaccines for Pandemic Influenza*. vol. 333, R. W. Compans and W. A. Orenstein, Eds., ed: Springer Berlin Heidelberg, 2009, pp. 43-82.

Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J-F, Dougherty B.A., Merrick J.M., "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd," *Science,* vol. 269, pp. 496-512, 1995.

Gahery H., Daniel N., Charmeteau B., Ourth L., Jackson A., Andrieu M., Choppin J., Salmon D., Pialoux G., Guillet J-G., "New CD4+ and CD8+ T cell responses induced in chronically HIV type-1-infected patients after immunizations with an HIV type 1 lipopeptide vaccine," *AIDS Research & Human Retroviruses,* vol. 22, pp. 684-694, 2006.

Galán J.E., "Bacterial toxins and the immune system show me the in vivo targets," *The Journal of experimental medicine,* vol. 201, pp. 321-323, 2005.

Gan W., Zhao G., Xu H., Wu W., Du W., Huang J., Yu X., Hu X., "Reverse vaccinology approach identify an *Echinococcus granulosus* tegumental membrane protein enolase as vaccine candidate," *Parasitology research,* vol. 106, pp. 873-882, 2010.

Gao J., Faraggi E., Zhou Y., Ruan J., Kurgan L., "BEST: improved prediction of B-cell epitopes from antigen sequences," *PloS one,* vol. 7, p. e40104, 2012.

Garmory H.S., Titball R.W., "ATP-binding cassette transporters are targets for the development of antibacterial vaccines and therapies," *Infection and immunity,* vol. 72, pp. 6757-6763, 2004.

Gay C., Zuerner R., Bannantine J., Lillehoj H., Zhu J., Green R., Pastoret P., "Genomics and vaccine development," *Revue scientifique et technique (International Office of Epizootics),* vol. 26, p. 49, 2007.

Gebriel A., Subramaniam G., Sekaran S., "The detection and characterization of pathogenic Leptospira and the use of OMPs as potential antigens and immunogens," *Trop Biomed,* vol. 23, pp. 194-207, 2006.

Geison G.L., "Pasteur's work on rabies: reexamining the ethical issues," *Hastings Center Report,* vol. 8, pp. 26-33, 1978.

Gherardi M.M., Esteban M., "Recombinant poxviruses as mucosal vaccine vectors," *Journal of general virology,* vol. 86, pp. 2925-2936, 2005.

Giefing C., Meinke A.L., Hanner M., Henics T., Minh D.B., Gelbmann D., Lundberg U., Senn B.M., Schunn M., Habel A., "Discovery of a novel class of highly conserved vaccine antigens using genomic scale antigenic fingerprinting of pneumococcus with human antibodies," *The Journal of experimental medicine,* vol. 205, pp. 117-131, 2008.

Girard M.P., Cherian T., Pervikov Y., Kieny M.P., "A review of vaccine research and development: human acute respiratory infections," *Vaccine,* vol. 23, pp. 5708-5724, 2005.

Giuliani M.M., Adu-Bobie J., Comanducci M., Aricò B., Savino S., Santini L., Brunelli B., Bambini S., Biolchi A., Capecchi B., "A universal vaccine for serogroup B

meningococcus," *Proceedings of the National Academy of Sciences,* vol. 103, pp. 10834-10839, 2006.

Glover D.T., Hollingshead S.K., Briles D.E., "*Streptococcus pneumoniae* surface protein PcpA elicits protection against lung infection and fatal sepsis," *Infection and immunity,* vol. 76, pp. 2767-2776, 2008.

Gomez G., Pei J., Mwangi W., Adams L.G., Rice-Ficht A., Ficht T.A., "Immunogenic and Invasive Properties of *Brucella melitensis* 16M Outer Membrane Protein Vaccine Candidates Identified via a Reverse Vaccinology Approach," *PloS one,* vol. 8, p. e59751, 2013.

Gong Y., Xu W., Cui Y., Zhang X., Yao R., Li D., Wang H., He Y., Cao J., Yin Y., "Immunization with a ZmpB-based protein vaccine could protect against pneumococcal diseases in mice," *Infection and immunity,* vol. 79, pp. 867-878, 2011.

Goodman A.G., Heinen P.P., Guerra S., Vijayan A., Sorzano C.O.S., Gomez C.E., Esteban M., "A human multi-epitope recombinant vaccinia virus as a universal T cell vaccine candidate against influenza virus," *PloS one,* vol. 6, p. e25938, 2011.

Gowthaman U., Agrewala J.N., "In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion," *Journal of proteome research,* vol. 7, pp. 154-163, 2007.

Grimes G.R., Moodie S., Beattie J.S., Craigon M., Dickinson P., Forster T., Livingston A.D., Mewissen M., Robertson K.A., Ross A.J., "GPX-Macrophage Expression Atlas: A database for expression profiles of macrophages challenged with a variety of pro-inflammatory, anti-inflammatory, benign and pathogen insults," *BMC Genomics,* vol. 6, p. 178, 2005.

Gross C.P., Sepkowitz K.A., "The myth of the medical breakthrough: smallpox, vaccination, and Jenner reconsidered," *International journal of infectious diseases,* vol. 3, pp. 54-60, 1998.

Grossman Z., Paul W.E., "Autoreactivity, dynamic tuning and selectivity," *Current opinion in immunology,* vol. 13, pp. 687-698, 2001.

Günther S., Hempel D., Dunkel M., Rother K., Preissner R., "SuperHapten: a comprehensive database for small immunogenic compounds," *Nucleic acids research,* vol. 35, pp. D906-D910, 2007.

Hagan E.C., Mobley H.L., "Uropathogenic *Escherichia coli* outer membrane antigens expressed during urinary tract infection," *Infection and immunity,* vol. 75, pp. 3941-3949, 2007.

Hamel J., Charland N., Pineau I., Ouellet C., Rioux S., Martin D., Brodeur B.R., "Prevention of pneumococcal disease in mice immunized with conserved surface-accessible proteins," *Infection and immunity,* vol. 72, pp. 2659-2670, 2004.

Hatfaludi T., Al-Hasani K., Gong L., Boyce J.D., Ford M., Wilkie I.W., Quinsey N., Dunstone M.A., Hoke D.E., Adler B., "Screening of 71 P. multocida proteins for protective efficacy in a fowl cholera infection model and characterization of the protective antigen PlpE," *PloS one,* vol. 7, p. e39973, 2012.

Hayes C.N., Diez D., Joannin N., Honda W., Kanehisa M., Wahlgren M., Wheelock C.E., Goto S., "varDB: a pathogen-specific sequence database of protein families involved in antigenic variation," *Bioinformatics,* vol. 24, pp. 2564-2565, 2008.

Häyrinen J., Jennings H., Raff H.V., Rougon G., Hanai N., Gerardy-Schahn R., Finne J., "Antibodies to polysialic acid and its N-propyl derivative: binding properties and interaction with human embryonal brain glycopeptides," *Journal of Infectious Diseases,* vol. 171, pp. 1481-1490, 1995.

He Y., Xiang Z., Mobley H.L., "Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development," *BioMed Research International,* vol. 2010, 2010.

Henderson B., Nair S., Pallas J., Williams M.A., "Fibronectin: a multidomain host adhesin targeted by bacterial fibronectin-binding proteins," *FEMS microbiology reviews,* vol. 35, pp. 147-200, 2011.

Hilleman M.R., "Past, present, and future of measles, mumps, and rubella virus vaccines," *Pediatrics,* vol. 90, pp. 149-153, 1992.

Hong M., Ahn J., Yoo S., Hong J., Lee E., Yoon I., Jung J-k., Lee H., "Identification of novel immunogenic proteins in pathogenic *Haemophilus parasuis* based on genome sequence analysis," *Veterinary microbiology,* vol. 148, pp. 89-92, 2011.

Hoof I., Peters B., Sidney J., Pedersen L.E., Sette A., Lund O., Buus S., Nielsen M., "NetMHCpan, a method for MHC class I binding prediction beyond humans," *Immunogenetics,* vol. 61, pp. 1-13, 2009.

Huang J., Honda W., "CED: a conformational epitope database," *BMC Immunology,* vol. 7, p. 7, 2006.

Iwai L.K., Juliano M.A., Juliano L., Kalil J., Cunha-Neto E., "T-cell molecular mimicry in Chagas disease: identification and partial structural analysis of multiple cross-reactive epitopes between *Trypanosoma cruzi* B13 and cardiac myosin heavy chain," *Journal of autoimmunity,* vol. 24, pp. 111-117, 2005.

Jackson D.A, Symons R.H, Berg P., "Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*," *Proceedings of the National Academy of Sciences,* vol. 69, pp. 2904-2909, 1972.

Jefferson T., Demicheli U., Pratt M., "Vaccines for preventing plague," *Cochrane database of systematic reviews,* 1998.

Jeong, S. H., Qiao, M., Nascimbeni, M., Hu, Z., Rehermann, B., Murthy, K., Liang, T. J., "Immunization with hepatitis C virus-like particles induces humoral and cellular immune responses in nonhuman primates," *Journal of virology*, vol. 78, p. 6995-7003, 2004.

John L., John G.J., Kholia T., "A reverse vaccinology approach for the identification of potential vaccine candidates from Leishmania spp," *Applied biochemistry and biotechnology,* vol. 167, pp. 1340-1350, 2012.

Johnston M.I., Fauci A.S., "An HIV vaccine—challenges and prospects," *New England Journal of Medicine,* vol. 359, pp. 888-890, 2008.

Karosiene E., Rasmussen M., Blicher T., Lund O., Buus S., Nielsen M., "NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ," *Immunogenetics,* pp. 1-14, 2013.

Kaur K., Sullivan M., Wilson P.C., "Targeting B cell responses in universal influenza vaccine design," *Trends in immunology,* vol. 32, pp. 524-531, 2011.

Khan J.M., Cheruku H.R., Tong J.C., Ranganathan S., "MPID-T2: a database for sequence–structure–function analyses of pMHC and TR/pMHC structures," *Bioinformatics,* vol. 27, pp. 1192-1193, 2011.

Kimman T., "Risks connected with the use of conventional and genetically engineerd vaccines," *Veterinary Quarterly,* vol. 14, pp. 110-118, 1992.

Kindsmüller, K., & Wagner, R., "Synthetic biology," *Human vaccines*, vol. 7, pp. 658-662, 2011.

Kirkwood C.D., "Genetic and antigenic diversity of human rotaviruses: potential impact on vaccination programs," *Journal of Infectious Diseases,* vol. 202, pp. S43-S48, 2010.

Ko J., Splitter G.A., "Molecular host-pathogen interaction in brucellosis: current understanding and future approaches to vaccine development for mice and humans," *Clinical microbiology reviews,* vol. 16, pp. 65-78, 2003.

Koff W.C., Burton D.R., Johnson P.R., Walker B.D., King C.R., Nabel G.J., Ahmed R., Bhan M.K., Plotkin S.A., "Accelerating Next-Generation Vaccine Development for Global Disease Prevention," *Science,* vol. 340, 2013.

Kopecky-Bromberg S.A., Palese P., "Recombinant vectors as influenza vaccines," in *Vaccines for Pandemic Influenza*, ed: Springer, 2009, pp. 243-267.

Košík I., Krejnusová I., Práznovská M., Poláková K., Russ G., "A DNA vaccine expressing PB1 protein of influenza A virus protects mice against virus infection," *Archives of virology,* vol. 157, pp. 811-817, 2012.

Kringelum J.V., Lundegaard C., Lund O., Nielsen M., "Reliable B cell epitope predictions: impacts of method development and improved benchmarking," *PLoS computational biology,* vol. 8, p. e1002829, 2012.

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., "Initial sequencing and analysis of the human genome," *Nature,* vol. 409, pp. 860-921, 2001.

Langermann S., Möllby R., Burlein J.E., Palaszynski S.R., Auguste C.G., DeFusco A., Strouse R., Schenerman M.A., Hultgren S.J., Pinkner J.S., "Vaccination with FimH Adhesin Protects Cynomolgus Monkeys from Colonization and Infection by Uropathogenic *Escherichia coli*," *Journal of Infectious Diseases,* vol. 181, pp. 774-778, 2000.

Larsen J.E., Lund O., Nielsen M., "Improved method for predicting linear B-cell epitopes," *Immunome research,* vol. 2, p. 2, 2006.

Larsen M.V., Lundegaard C., Lamberth K., Buus S., Lund O., Nielsen M., "Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction," *BMC Bioinformatics,* vol. 8, p. 424, 2007.

Lee J.S., Shin S.J., Collins M.T., Jung I.D., Jeong Y-I., Lee C-M., Shin Y.K., Kim D., Park Y-M., "*Mycobacterium avium* subsp. paratuberculosis fibronectin attachment protein activates dendritic cells and induces a Th1 polarization," *Infection and immunity,* vol. 77, pp. 2979-2988, 2009.

Lefranc M-P., Giudicelli V., Kaas Q., Duprat E., Jabado-Michaloud J., Scaviner D., Ginestoux C., Clement O., Chaume D., Lefranc G., "IMGT, the international ImMunoGeneTics information system®," *Nucleic acids research,* vol. 33, pp. D593-D597, 2005.

Liang S., Zheng D., Standley D., Yao B., Zacharias M., Zhang C., "EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results," *BMC Bioinformatics,* vol. 11, p. 381, 2010.

Liang S., Zheng D., Zhang C., Zacharias M., "Prediction of antigenic epitopes on protein surfaces by consensus scoring," *BMC Bioinformatics,* vol. 10, p. 302, 2009.

Liebenberg J., Pretorius A., Faber F., Collins N.E., Allsopp B.A., Van Kleef M., "Identification of *Ehrlichia ruminantium* proteins that activate cellular immune responses using a reverse vaccinology strategy," *Veterinary Immunology and Immunopathology,* vol. 145, pp. 340-349, 2012.

Liechti R., Gleizes A., Kuznetsov D., Bougueleret L., Le Mercier P., Bairoch A., Xenarios I., "OpenFluDB, a database for human and animal influenza virus," *Database: the journal of biological databases and curation,* vol. 2010, 2010.

Liu L., Cheng G., Wang C., Pan X., Cong Y., Pan Q., Wang J., Zheng F., Hu F., Tang J., "Identification and experimental verification of protective antigens against *Streptococcus suis* serotype 2 based on genome sequence analysis," *Current microbiology,* vol. 58, pp. 11-17, 2009.

Loosmore S.M., Yang Y-p., Oomen R., Shortreed J.M., Coleman D.C., Klein M.H., "The *Haemophilus influenzae* HtrA protein is a protective antigen," *Infection and immunity,* vol. 66, pp. 899-906, 1998.

Lundegaard C., Lund O., Nielsen M., "Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers," *Bioinformatics,* vol. 24, pp. 1397-1398, 2008.

Lynn D.J., Winsor G.L., Chan C., Richard N., Laird M.R., Barsky A., Gardy J.L., Roche F.M., Chan T.H., Shah N., "InnateDB: facilitating systems-level analyses of the mammalian innate immune response," *Molecular systems biology,* vol. 4, 2008.

Magnan C.N., Zeller M., Kayala M.A., Vigil A., Randall A., Felgner P.L., Baldi P., "High-throughput prediction of protein antigenicity using protein microarray data," *Bioinformatics,* vol. 26, pp. 2936-2943, 2010.

Manque P.A., Tenjo F., Woehlbier U., Lara A.M., Serrano M.G., Xu P., Alves J.M., Smeltz R.B., Conrad D.H., Buck G.A., "Identification and immunological characterization of three potential vaccinogens against Cryptosporidium species," *Clinical and Vaccine Immunology,* vol. 18, pp. 1796-1802, 2011.

McMurry J.A., Johansson B.E., De Groot A.S., "A call to cellular & humoral arms: enlisting cognate T cell help to develop broad-spectrum vaccines against influenza A," *Human vaccines,* vol. 4, pp. 148-157, 2008.

McSparron H., Blythe M.J., Zygouri C., Doytchinova I.A., Flower D.R., "JenPep: a novel computational information resource for immunobiology and vaccinology," *Journal of chemical information and computer sciences,* vol. 43, pp. 1276-1287, 2003.

Michel, M. L., & Tiollais, P., "Hepatitis B vaccines: protective efficacy and therapeutic potential," *Pathologie Biologie*, vol. 58, p. 288-295, 2010.

Montigiani S., Falugi F., Scarselli M., Finco O., Petracca R., Galli G., Mariani M., Manetti R., Agnusdei M., Cevenini R., "Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*," *Infection and immunity,* vol. 70, pp. 368-379, 2002.

Moriel D.G., Bertoldi I., Spagnuolo A., Marchi S., Rosini R., Nesta B., Pastorello I., Corea VAM., Torricelli G., Cartocci E., "Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic *Escherichia coli*," *Proceedings of the National Academy of Sciences,* vol. 107, pp. 9072-9077, 2010.

Morris S.K., Moss W.J., Halsey N., "*Haemophilus influenzae* type b conjugate vaccine use and effectiveness," *The Lancet infectious diseases,* vol. 8, pp. 435-443, 2008.

Muh H.C., Tong J.C., Tammi M.T., "AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins," *PloS one,* vol. 4, p. e5861, 2009.

Murray G.L., Lo M., Bulach D.M., Srikram A., Seemann T., Quinsey N.S., Sermswan R.W., Allen A., Adler B., "Evaluation of 238 antigens of *Leptospira borgpetersenii* serovar Hardjo for protection against kidney colonisation," *Vaccine,* 2012.

Nagy G., Pál T., "Strategies for the development of vaccines conferring broad-spectrum protection," *International Journal of Medical Microbiology,* vol. 298, pp. 379-395, 2008.

Nancy Y.Y., Wagner J.R., Laird M.R., Melli G., Rey S., Lo R., Dao P., Sahinalp S.C., Ester M., Foster L.J., "PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes," *Bioinformatics,* vol. 26, pp. 1608-1615, 2010.

Nielsen M., Lund O., "NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction," *BMC Bioinformatics,* vol. 10, p. 296, 2009.

Nobbs A.H., Lamont R.J., Jenkinson H.F., "Streptococcus adherence and colonization," *Microbiology and Molecular Biology Reviews,* vol. 73, pp. 407-450, 2009.

Normile D., "Driven to Extinction," *science,* vol. 319, pp. 1606-1609, March 21, 2008 2008.

Ogunniyi A.D., Woodrow M.C., Poolman J.T., Paton J.C., "Protection against *Streptococcus pneumoniae* elicited by immunization with pneumolysin and CbpA," *Infection and immunity,* vol. 69, pp. 5997-6003, 2001.

Oprea M., Antohe F., “Reverse-vaccinology strategy for designing T-cell epitope candidates for *Staphylococcus aureus* endocarditis vaccine,” *Biologicals,* 2013.

Osterholm M.T., Kelley N.S., Sommer A., Belongia E.A., “Efficacy and effectiveness of influenza vaccines: a systematic review and meta-analysis,” *The Lancet infectious diseases,* vol. 12, pp. 36-44, 2012.

Overweg K., Kerr A., Sluijter M., Jackson M., Mitchell T., De Jong A., De Groot R., Hermans P., “The Putative Proteinase Maturation Protein A of *Streptococcus pneumoniae* Is a Conserved Surface Protein with Potential To Elicit Protective Immune Responses,” *Infection and immunity,* vol. 68, pp. 4180-4188, 2000.

Palumbo R.N., Wang C., “Bacterial invasin: structure, function, and implication for targeted oral gene delivery,” *Current Drug Delivery,* vol. 3, pp. 47-53, 2006.

Pasteur L., “De l'attenuation du virus du cholera des poules,” *CR Acad. Sci. Paris,* vol. 91, pp. 673-680, 1880.

Peters B., Tong W., Sidney J., Sette A., Weng Z., “Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules,” *Bioinformatics,* vol. 19, pp. 1765-1772, 2003.

Petrovsky N., Schönbach C., Brusic V., “Bioinformatic strategies for better understanding of immune function,” *In silico biology,* vol. 3, pp. 411-416, 2003.

Pizza M., Scarlato V., Masignani V., Giuliani M.M., Arico B., Comanducci M., Jennings G.T., Baldi L., Bartolini E., Capecchi B., “Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing,” *science,* vol. 287, pp. 1816-1820, 2000.

Plotkin S.L., Plotkin S.A., “A short history of vaccination,” 2004.

Ponomarenko J., Bui H-H., Li W., Fusseder N., Bourne P.E., Sette A., Peters B., “ElliPro: a new structure-based tool for the prediction of antibody epitopes,” *BMC Bioinformatics,* vol. 9, p. 514, 2008.

Ponomarenko J.V., Bourne P.E., “Antibody-protein interactions: benchmark datasets and prediction tools evaluation,” *BMC Structural biology,* vol. 7, p. 64, 2007.

Poolman J.T., “Development of a meningococcal vaccine,” *Infectious agents and disease,* vol. 4, p. 13, 1995.

Potter A.A., Schryvers A.B., Ogunnariwo J.A., Hutchins W.A., Lo R.Y., Watts T., “Protective capacity of the *Pasteurella haemolytica* transferrin-binding proteins TbpA and TbpB in cattle,” *Microbial pathogenesis,* vol. 27, pp. 197-206, 1999.

Premendu P.M., "Structural Epitope Database (SEDB): A Web-based Database for the Epitope, and its Intermolecular Interaction Along with the Tertiary Structure Information," *Journal of Proteomics & Bioinformatics,* 2012.

Puck J.M., de Saint Basile G., Schwarz K., Fugmann S., Fischer R.E., "IL2RGbase: a database of γc-chain defects causing human X-SCID," *Immunology today,* vol. 17, pp. 507-511, 1996.

Punta M., Coggill P.C., Eberhardt R.Y., Mistry J., Tate J., Boursnell C., Pang N., Forslund K., Ceric G., Clements J., "The Pfam protein families database," *Nucleic acids research,* vol. 40, pp. D290-D301, 2012.

Quan, F. S., Huang, C., Compans, R. W., & Kang, S. M., "Virus-like particle vaccine induces protective immunity against homologous and heterologous strains of influenza virus," *Journal of virology*, vol. 81, p. 3514-3524, 2007.

Racz R., Chung M., Xiang Z., He Y., "Systematic annotation and analysis of "virmugens"—Virulence factors whose mutants can be used as live attenuated vaccines," *Vaccine,* 2012.

Rajapakse M., Zhang G.L., Srinivasan K.N., Schmidt B., Petrovsky N., Brusic V., "PREDNOD, a prediction server for peptide binding to the H-2g7 haplotype of the non-obese diabetic mouse," *Autoimmunity,* vol. 39, pp. 645-650, 2006.

Rammensee H-G., Bachmann J., Emmerich N.P.N., Bachor O.A., Stevanović S., "SYFPEITHI: database for MHC ligands and peptide motifs," *Immunogenetics,* vol. 50, pp. 213-219, 1999.

Ramos H.C., Rumbo M., Sirard J-C., "Bacterial flagellins: mediators of pathogenicity and host immune responses in mucosa," *Trends in microbiology,* vol. 12, pp. 509-517, 2004.

Rao V., Dhar N., Tyagi A., "Modulation of host immune responses by overexpression of immunodominant antigens of *Mycobacterium tuberculosis* in bacille Calmette–Guerin," *Scandinavian journal of immunology,* vol. 58, pp. 449-461, 2003.

Rappuoli R., "Bridging the knowledge gaps in vaccine design," *Nature biotechnology,* vol. 25, pp. 1361-1366, 2007.

Rappuoli R., "Reverse vaccinology," *Current opinion in microbiology,* vol. 3, pp. 445-450, 2000.

Rappuoli R., Pizza M., Douce G., Dougan G., "8. New Vaccines against Bacterial Toxins," *Advances in Experimental Medicine and Biology,* vol. 397, pp. 55-60, 1996.

Ratledge C., Dover L.G., "Iron metabolism in pathogenic bacteria," *Annual Reviews in Microbiology,* vol. 54, pp. 881-941, 2000.

Reche P.A., Glutting J-P., Reinherz E.L., "Prediction of MHC class I binding peptides using profile motifs," *Human immunology,* vol. 63, pp. 701-709, 2002.

Regoes R.R., Bonhoeffer S., "Emergence of drug-resistant influenza virus: population dynamical considerations," *science,* vol. 312, pp. 389-391, 2006.

Retter I., Althaus H.H., Münch R., Müller W., "VBASE2, an integrative V gene database," *Nucleic acids research,* vol. 33, pp. D671-D674, 2005.

Riedel S., "Edward Jenner and the history of smallpox and vaccination," *Proceedings (Baylor University. Medical Center),* vol. 18, p. 21, 2005.

Rimmelzwaan G., Boon A., Voeten J., Berkhoff E., Fouchier R., Osterhaus A., "Sequence variation in the influenza A virus nucleoprotein associated with escape from cytotoxic T lymphocytes," *Virus research,* vol. 103, pp. 97-100, 2004.

Robinson J., Halliwell J.A., McWilliam H., Lopez R., Marsh S.G., "IPD—the Immuno Polymorphism Database," *Nucleic acids research,* vol. 41, pp. D1234-D1240, 2013.

Robinson J., Waller M.J., Parham P., de Groot N., Bontrop R., Kennedy L.J., Stoehr P., Marsh S.G., "IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex," *Nucleic acids research,* vol. 31, pp. 311-314, 2003.

Robinson J., Waller M.J., Stoehr P., Marsh S.G., "IPD—the immuno polymorphism database," *Nucleic acids research,* vol. 33, pp. D523-D526, 2005.

Roldão, A., Mellado, M. C. M., Castilho, L. R., Carrondo, M. J., & Alves, P. M., "Virus-like particles in vaccine development," *Expert review of vaccines*, vol. 9, pp. 1149-1176, 2010.

Rosenow C., Ryan P., Weiser J.N., Johnson S., Fontan P., Ortqvist A., Masure H.R., "Contribution of novel choline-binding proteins to adherence, colonization and immunogenicity of *Streptococcus pneumoniae*," *Molecular microbiology,* vol. 25, pp. 819-829, 1997.

Ross B.C., Czajkowski L., Hocking D., Margetts M., Webb E., Rothel L., Patterson M., Agius C., Camuglia S., Reynolds E., "Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis*" *Vaccine,* vol. 19, pp. 4135-4142, 2001.

Rubinstein N.D., Mayrose I., Martz E., Pupko T., "Epitopia: a web-server for predicting B-cell epitopes," *BMC Bioinformatics,* vol. 10, p. 287, 2009.

Sadilkova L., Nepereny J., Vrzal V., Sebo P., Osicka R., "Type IV fimbrial subunit protein ApfA contributes to protection against porcine pleuropneumonia," *Veterinary research,* vol. 43, pp. 1-12, 2012.

Saha S., Bhasin M., Raghava G.P., "Bcipep: a database of B-cell epitopes," *BMC Genomics,* vol. 6, p. 79, 2005.

Saha S., Raghava G., "AlgPred: prediction of allergenic proteins and mapping of IgE epitopes," *Nucleic acids research,* vol. 34, pp. W202-W209, 2006.

Saha S., Raghava G., "BcePred: Prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties," in *Artificial Immune Systems*, ed: Springer, 2004, pp. 197-204.

Saha S., Raghava G., "Prediction of continuous B-cell epitopes in an antigen using recurrent neural network," *Proteins: Structure, Function, and Bioinformatics,* vol. 65, pp. 40-48, 2006.

Sauvage E., Kerff F., Terrak M., Ayala J.A., Charlier P., "The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis," *FEMS microbiology reviews,* vol. 32, pp. 234-258, 2008.

Schorey J.S., Holsti M.A., Ratliff T.L., Allen P.M., Brown E.J., "Characterization of the fibronectin-attachment protein of *Mycobacterium avium* reveals a fibronectin-binding motif conserved among mycobacteria," *Molecular microbiology,* vol. 21, pp. 321-329, 1996.

Sedova E., Shcherbinin D., Migunov A., Smirnov I.A., Logunov D.I., Shmarov M., Tsybalova L., Naroditskiĭ B., Kiselev O., Gintsburg A., "Recombinant Influenza Vaccines," *Acta naturae,* vol. 4, p. 17, 2012.

Sidney J., Peters B., Frahm N., Brander C., Sette A., "HLA class I supertypes: a revised and updated classification," *BMC Immunology,* vol. 9, p. 1, 2008.

Singh H., Ansari H.R., Raghava G.P., "Improved Method for Linear B-Cell Epitope Prediction Using Antigen's Primary Sequence," *PloS one,* vol. 8, p. e62216, 2013.

Singh H., Raghava G., "ProPred: prediction of HLA-DR binding sites," *Bioinformatics,* vol. 17, pp. 1236-1237, 2001.

Singh H., Raghava G., "ProPred1: prediction of promiscuous MHC Class-I binding sites," *Bioinformatics,* vol. 19, pp. 1009-1014, 2003.

Singh M.K., Srivastava S., Raghava G., Varshney G.C., "HaptenDB: a comprehensive database of haptens, carrier proteins and anti-hapten antibodies," *Bioinformatics,* vol. 22, pp. 253-255, 2006.

Sinha A., Levine O., Knoll M.D., Muhib F., Lieu T.A., "Cost-effectiveness of pneumococcal conjugate vaccination in the prevention of child mortality: an international economic analysis," *The Lancet,* vol. 369, pp. 389-396, 2007.

Squires R.B., Noronha J., Hunt V., García-Sastre A., Macken C., Baumgarth N., Suarez D., Pickett B.E., Zhang Y., Larsen C.N., "Influenza Research Database: an integrated bioinformatics resource for influenza research and surveillance," *Influenza and other respiratory viruses,* vol. 6, pp. 404-416, 2012.

Sun J., Wu D., Xu T., Wang X., Xu X., Tao L., Li Y., Cao Z-W., "SEPPA: a computational server for spatial epitope prediction of protein antigens," *Nucleic acids research,* vol. 37, pp. W612-W616, 2009.

Sweredoski M.J., Baldi P., "COBEpro: a novel system for predicting continuous B-cell epitopes," *Protein Engineering Design and Selection,* vol. 22, pp. 113-120, 2009.

Sweredoski M.J., Baldi P., "PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure," *Bioinformatics,* vol. 24, pp. 1459-1460, 2008.

Sylte M.J., Suarez D.L., "Influenza neuraminidase as a vaccine antigen," in *Vaccines for Pandemic Influenza*, ed: Springer, 2009, pp. 227-241.

Talkington D.F., Brown B.G., Tharpe J.A., Koenig A., Russell H., "Protection of mice against fatal pneumococcal challenge by immunization with pneumococcal surface adhesin A (PsaA)," *Microbial pathogenesis,* vol. 21, pp. 17-22, 1996.

Tan P.T., Khan A.M., August J.T., "Highly conserved influenza A sequences as T cell epitopes-based vaccine targets to address the viral variability," *Human vaccines,* vol. 7, pp. 402-409, 2011.

Tang C., Holden D., "Pathogen virulence genes–implications for vaccines and drug therapy," *British medical bulletin,* vol. 55, pp. 387-400, 1999.

Tarca A.L., Carey V.J., Chen X-w., Romero R., Drăghici S., "Machine learning and its applications to biology," *PLoS computational biology,* vol. 3, p. e116, 2007.

Theiler M., Smith H.H., "The use of yellow fever virus modified by in vitro cultivation for human immunization," *The Journal of experimental medicine,* vol. 65, pp. 787-800, 1937.

Tong H., Li D., Chen S., Long J., DeMaria T., "Immunization with recombinant *Streptococcus pneumoniae* neuraminidase NanA protects chinchillas against nasopharyngeal colonization," *Infection and immunity,* vol. 73, pp. 7775-7778, 2005.

Tong J.C., Song C.M., Tan P.T.J., Ren E.C., Sinha A.A., "BEID: Database for sequence-structure-function information on antigen-antibody interactions," *Bioinformation,* vol. 3, p. 58, 2008.

Tung C-W., Ho S-Y., "POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties," *Bioinformatics,* vol. 23, pp. 942-949, 2007.

Turbyfill K.R., Kaminski R.W., Oaks E.V., "Immunogenicity and efficacy of highly purified invasin complex vaccine from *Shigella flexneri* 2a," *Vaccine,* vol. 26, pp. 1353-1364, 2008.

Tusnady G.E., Simon I., "The HMMTOP transmembrane topology prediction server," *Bioinformatics,* vol. 17, pp. 849-850, 2001.

Umamaheswari A., Pradhan D., Hemanthkumar M., "Computer aided subunit vaccine design against pathogenic Leptospira serovars," *Interdisciplinary Sciences: Computational Life Sciences,* vol. 4, pp. 38-45, 2012.

Van den Eynde B.J., van der Bruggen P., "T cell defined tumor antigens," *Current opinion in immunology,* vol. 9, pp. 684-693, 1997.

Vemulapalli R., He Y., Cravero S., Sriranganathan N., Boyle S.M., Schurig G.G., "Overexpression of Protective Antigen as a Novel Approach To Enhance Vaccine Efficacy of *Brucella abortus* Strain RB51," *Infection and immunity,* vol. 68, pp. 3286-3289, 2000.

Vita R., Zarebski L., Greenbaum J.A., Emami H., Hoof I., Salimi N., Damle R., Sette A., Peters B., "The immune epitope database 2.0," *Nucleic acids research,* vol. 38, pp. D854-D862, 2010.

Vivona S., Bernante F., Filippini F., "NERVE: new enhanced reverse vaccinology environment," *BMC Biotechnology,* vol. 6, p. 35, 2006.

Wan J., Liu W., Xu Q., Ren Y., Flower D.R., Li T., "SVRMHC prediction server for MHC-binding peptides," *BMC Bioinformatics,* vol. 7, p. 463, 2006.

Wang X., Zhao H., Xu Q., Jin W., Liu C., Zhang H., Huang Z., Zhang X., Zhang Y., Xin D., "HPtaa database-potential target genes for clinical diagnosis and immunotherapy of human carcinoma," *Nucleic acids research,* vol. 34, pp. D607-D612, 2006.

Wang Y., Zhou L., Shi H., Xu H., Yao H., Xi X.G., Toyoda T., Wang X., Wang T., "Monoclonal antibody recognizing SLLTEVET epitope of M2 protein potently inhibited the replication of influenza A viruses in MDCK cells," *Biochemical and biophysical research communications,* vol. 385, pp. 118-122, 2009.

Warfield, K. L., & Aman, M. J., "Advances in virus-like particle vaccines for filoviruses," *Journal of Infectious Diseases*, vol. 204(suppl 3), p. S1053-S1059, 2011.

Wee L., Lim S.J., Ng L., Tong J.C., "Immunoinformatics: how in silico methods are re-shaping the investigation of peptide immune specificity," *Frontiers in bioscience (Elite edition),* vol. 4, p. 311, 2012.

Wee L.J., Simarmata D., Kam Y-W., Ng L.F., Tong J.C., "SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction," *BMC Genomics,* vol. 11, p. S21, 2010.

Weiss M.M., Weiss P.D., Weiss J.B., "Anthrax vaccine and public health policy," *American journal of public health,* vol. 97, pp. 1945-1951, 2007.

Wizemann T.M., Adamou J.E., Langermann S., "Adhesins as targets for vaccine development," *Emerging infectious diseases,* vol. 5, p. 395, 1999.

Wizemann T.M., Heinrichs J.H., Adamou J.E., Erwin A.L., Kunsch C., Choi G.H., Barash S.C., Rosen C.A., Masure H.R., Tuomanen E., "Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection," *Infection and immunity,* vol. 69, pp. 1593-1598, 2001.

Xiang Z., Todd T., Ku K.P., Kovacic B.L., Larson C.B., Chen F., Hodges A.P., Tian Y., Olenzek E.A., Zhao B., "VIOLIN: vaccine investigation and online information network," *Nucleic acids research,* vol. 36, pp. D923-D928, 2008.

Yamamoto M., McDaniel L.S., Kawabata K., Briles D.E., Jackson R.J., McGhee J.R., Kiyono H., "Oral immunization with PspA elicits protective humoral immunity against *Streptococcus pneumoniae* infection," *Infection and immunity,* vol. 65, pp. 640-644, 1997.

Yang B., Sayers S., Xiang Z., He Y., "Protegen: a web-based protective antigen database and analysis system," *Nucleic acids research,* vol. 39, pp. D1073-D1078, 2011.

Yang H-L., Zhu Y-Z., Qin J-H., He P., Jiang X-C., Zhao G-P., Guo X-K., "In silico and microarray-based genomic approaches to identifying potential vaccine candidates against *Leptospira interrogans*," *BMC Genomics,* vol. 7, p. 293, 2006.

Yang S., Lee J-Y., Lee J.S., Mitchell W.P., Oh H-B., Kang C., Kim K.H., "Influenza sequence and epitope database," *Nucleic acids research,* vol. 37, pp. D423-D430, 2009.

Zarantonelli M.L., Antignac A., Lancellotti M., Guiyoule A., Alonso J-M., Taha M-K., "Immunogenicity of meningococcal PBP2 during natural infection and protective activity of anti-PBP2 antibodies against meningococcal bacteraemia in mice," *Journal of Antimicrobial Chemotherapy,* vol. 57, pp. 924-930, 2006.

Zhang H., Wang P., Papangelopoulos N., Xu Y., Sette A., Bourne P.E., Lund O., Ponomarenko J., Nielsen M., Peters B., "Limitations of Ab initio predictions of peptide binding to MHC class II molecules," *PloS one,* vol. 5, p. e9272, 2010.

Zhang M., Wu H., Li X., Yang M., Chen T., Wang Q., Liu Q., Zhang Y., " *Edwardsiella tarda* flagellar protein FlgD: A protective immunogen against edwardsiellosis," *Vaccine,* vol. 30, pp. 3849-3856, 2012.

Zhang W., Niu Y., Xiong Y., Zhao M., Yu R., Liu J., "Computational Prediction of Conformational B-Cell Epitopes from Antigen Primary Structures by Ensemble Learning," *PloS one,* vol. 7, p. e43575, 2012.

Zhang W., Xiong Y., Zhao M., Zou H., Ye X., Liu J., "Prediction of conformational B-cell epitopes from 3D structures by random forests with a distance-based feature," *BMC Bioinformatics,* vol. 12, p. 341, 2011.

Zou L., Wang J., Huang B., Xie M., Li A., "A solute-binding protein for iron transport in *Streptococcus iniae*," *BMC Microbiology,* vol. 10, p. 309, 2010.