

CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD



Reverse Vaccinology Approach
for the *Streptococcus gordonii* to
Identify Potential Antigenic
Determinants

by

Aneeqa Abid

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Health and Life Sciences

Department of Bioinformatics and Biosciences

2022

Copyright © 2022 by Aneeqa Abid

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.

I dedicate this thesis to all the great people came in my life specially my beloved Parents and my Supervisor who encouraged me to stand out in this world with nobility and motivated me to step ahead without any fear.



CERTIFICATE OF APPROVAL

Reverse Vaccinology Approach for the *Streptococcus gordonii* to Identify Potential Antigenic Determinants

by

Aneeqa Abid

(MBS203010)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Samiullah Khan	QAU, Islamabad
(b)	Internal Examiner	Dr. Erum Dilshad	CUST, Islamabad
(c)	Supervisor	Dr. Syeda Marriam Bakhtiar	CUST, Islamabad

Dr. Syeda Marriam Bakhtiar

Thesis Supervisor

November, 2022

Dr. Syeda Marriam Bakhtiar
Head
Dept. of Bioinfo. and Biosciences
November, 2022

Dr. Sahar Fazal
Dean
Faculty of Health and Life Sciences
November, 2022

Author's Declaration

I, **Aneeqa Abid** hereby state that my MS thesis titled “**Reverse Vaccinology Approach for the *Streptococcus gordonii* to Identify Potential Antigenic Determinants**” is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Aneeqa Abid)

Registration No: MBS203010

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled “**Reverse Vaccinology Approach for the *Streptococcus gordonii* to Identify Potential Antigenic Determinants**” is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Aneeqa Abid)

Registration No: MBS203010

Acknowledgement

In the name of Allah Almighty, the most Gracious and the most Merciful, first and foremost, I am grateful to Almighty ALLAH for giving me the strength, aptitude, ability, knowledge and opportunity to undertake this study and complete it satisfactorily. Indeed, all praises are due to Him and His Holy Prophet Muhammad (PBUH). Secondly, I would like to express my gratitude to Capital University of Science and Technology (CUST) Islamabad my Alma for providing me the opportunity to do MS Biosciences and achieve my goal of pursuing higher studies and I want to acknowledge the efforts of my thesis supervisor **Dr. Syeda Marriam Bakhtiar**, Department of Biosciences and Bioinformatics, whose worthy guidance, encouragement, and professional attitude is appreciable in completing this dissertation. Also, I would like to thank my respected co-supervisor **Dr. Syed Babar Jamal Bacha**, National University of Medical Sciences who supported, motivated, guided and facilitated to carry out this research work. I thankfully acknowledge the support and inspiration that I received from **Dr. Muhammad Faheem** and appreciate his continues help during my research work.

I want to acknowledge Dean of Faculty of Health and Life Sciences, **Dr. Sahar Fazal**, and head of Department of Bioinformatics and Biosciences, Dr. Syeda Marriam Bakhtiar, for giving me the opportunity to pursue MS with thesis. I owe a great deal of appreciation and gratitude to all the Faculty members, Dr. Shaukat Iqbal Malik, Dr. Erum Dilshad, Dr. Sohail Ahmed Jan and Dr. Arshia Amin Butt for the support and inspiration. I thankfully acknowledge the support and inspiration that I received time to time from all my friends and seniors. I am deeply grateful to my parents and my siblings for their support, appreciation, encouragement and keen interest in my achievements without which I was unable to do anything. I really thank my family for their moral support, love, care, and prayers throughout the research work.

(Aneeqa Abid)

Abstract

Streptococcus gordonii is member of the group of viridians bacteria (α -hemolytic-sanguinis). It is Gram positive, mutualistic bacterium found in the human body, involving the oral cavity, upper respiratory tract, skin, and intestines. It is also opportunistic pathogen that can cause various diseases and infections including infective endocarditis and apical periodontitis. *S. gordonii* enters the inside of tooth i.e., root canal and vessels of blood, after which they interact with numerous immune (leukocytes) and non-immune cells of the body. Antibiotic treatment is available against infection caused by *S. gordonii*. But antibiotic treatment is not long lasting and infection can be occurred again. Vaccine has long lasting impact on the organism to prevent infectious diseases. It is a preventive way to stop the onset of disease. So by developing vaccine, one can develop immunity against *S. gordonii* and prevent infections and diseases caused by it like infective endocarditis (IE). 13 complete strains out of 91 strains of *S.gordonii* were selected. Out of 2835 genes, 1225 were core genes (pan-genomics). From 1255 core genes, 643 were identified as non-homologous proteins (Subtractive genomics). A total of 20 essential proteins were predicted from non-homologous proteins. The proteins collected for this study were surface proreins and out of 20 essential proteins, only 5 were identified. The antigenic proteins were selected and vaccine was designed on the basis of selected B and T-cell epitopes of the 2 antigenic proteins with the help of linkers and adjuvant. The designed vaccine was docked against TLR-2. The TLR-2 plays important role in human immune system. Vaccine I with adjuvant shows higher interactions. The expression of the protein was determined using in-silico gene cloning.

Keywords: Streptococcus gordonii, Infective endocarditis, Pan-genomics, Subtractive - genomics, Docking, TLR2

Contents

Author's Declaration	iv
Plagiarism Undertaking	v
Acknowledgement	vi
Abstract	vii
List of Figures	xii
List of Tables	xiv
Abbreviations	xv
1 Introduction	1
1.1 Aim and Objectives	3
2 Literature Review	5
2.1 Streptococcus	6
2.2 Oral Streptococci	6
2.2.1 Taxonomy of Oral Streptococci	6
2.3 Viridans Streptococci	8
2.4 <i>Streptococcus gordonii</i>	8
2.4.1 Pathogenicity	10
2.5 Infective Endocarditis	12
2.5.1 Prevalence of Infective Endocarditis in Pakistan	13
2.5.2 Antibiotic Treatment	14
2.5.3 Challenges in IE	14
2.6 Vaccinology	14
2.6.1 Types of Vaccine	15
2.7 Immune Response	15
2.8 Reverse Vaccinology	17
2.9 Pan-Genome	18
2.10 Scope of Study	18
2.11 Impact on Society	19
2.12 Gap Analysis	19

2.13	Research Question	19
3	Materials and Methods	21
3.1	Vaccine Target Prediction	21
3.1.1	Sequence Retrieval and Genome Selection	21
3.1.2	Identification of Core Genome	21
3.1.3	Homologous and Non-homologous Protein Identification by using Subtractive Genomic Analysis	22
3.1.4	Identification of Essential genes	22
3.2	Collection of Surface Proteins	22
3.2.1	Sequence Retrieval of Proteins	23
3.2.2	Screening Antigenicity and Allergenicity of <i>S. gordonii</i> Proteins and Protein Selection	23
3.2.3	Structure Analysis by Screening of Physio-chemical Properties of Selected Proteins	23
3.2.4	Function and Pathway Analysis of Selected Protein	24
3.3	Epitope Prediction	24
3.4	Prediction of B-Cell Epitopes (LBL) via Immunoinformatics	24
3.4.1	Antigenicity, Allergenicity, Toxicity, Mol. Weight and Sub-Cellular Localization Profiling of Epitopes	25
3.5	Prediction of T-cell Epitope via Immunoinformatics	25
3.5.1	MHC-I Binding Epitopes (Cytotoxic T-lymphocytes - CTL)	25
3.5.2	MHC-II Restricted epitopes (Helper T-cell Lymphocytes (HTL) Epitopes)	26
3.5.3	Antigenicity, Allergenicity, Toxicity and Sub-Cellular Localization Profiling of Epitopes	26
3.5.4	Determination of Physicochemical Properties and other Vital Features	27
3.6	Epitope Conservation Analysis	27
3.7	Population Coverage Prediction of T-Cell Epitopes	27
3.8	Construction of Multivalent Vaccine Design	28
3.8.1	Antigenicity and Allergenicity Profiling of Vaccine Construct	29
3.8.2	Physicochemical Properties and Solubility Prediction of Vaccine Construct	29
3.9	Structure Prediction, Refinement and Validation of Multi-epitope Vaccine	29
3.9.1	Prediction of Secondary Structure	30
3.9.2	Prediction of Vaccine 3D Structure	30
3.9.3	Refinement of Vaccine 3D Structure	30
3.9.4	Validation of Vaccine 3D Structure	30
3.9.5	Proteasomal Cleavages Prediction of MHC-ligands	30
3.10	Molecular Docking with TLR2	31
3.11	Immune-simulation	31
3.12	Gene Cloning	32
3.12.1	Sequence Translation of Vaccine Protein and Codon Adaptation	32

3.12.2	In-silico Cloning through SnapGene	32
3.13	Overview of Methodology	32
4	Results and Discussion	34
4.1	Vaccine Target Prediction	34
4.1.1	Sequence Retrieval and Genome Selection	34
4.1.2	Identification of Core Genome by Pan-genome Approach	34
4.1.3	Homologous and Non-homologous Protein Identification by using Subtractive Genomic Analysis	35
4.1.4	Identification of Essential Genes	35
4.2	Collection of Surface Proteins	35
4.2.1	Complete Sequence Retrieval of Proteins	36
4.2.2	Screening Antigenicity and Allergicity <i>S. gordonii</i> Proteins and Selection of Protein	37
4.2.3	Structure Analysis by Screening of Physio-chemical Properties of Selected Proteins	38
4.2.4	Function and Pathway Analysis of Selected Protein	40
4.3	B-cell Epitope Prediction and Selection	40
4.3.1	Antigenicity, Allergenicity, Molecular Weight and Sub-cellular Localization Profiling of Epitopes	41
4.4	T-cell Epitope Selection	42
4.4.1	Screening and Selection of MHC-I Restricted Epitopes	42
4.4.1.1	Antigenicity, Allergenicity, Molecular Weight and Sub-cellular Localization Profiling	45
4.4.1.2	Physiochemical Properties and Vital Features Determination	46
4.4.2	Screening of MHC-II Restricted Epitopes	49
4.4.2.1	Antigenicity, Allergenicity, Molecular Weight and Sub-cellular Localization Profiling	52
4.4.2.2	Determination of Physiochemical properties and Vital Features	53
4.5	Epitope Conservation Analysis	56
4.6	Population Coverage of Selected T-Cell Epitopes	57
4.7	Multivalent Vaccine Design Construction	60
4.7.1	Antigenicity and Allergenicity Prediction of Multivalent Vaccine Construct	62
4.7.2	Physicochemical Properties and Solubility Determination of Primary Structure of Vaccine Construct	62
4.8	Structure Prediction, Refinement and Validation of Multi-epitope Vaccine	64
4.8.1	Secondary Structure Prediction	64
4.8.2	Tertiary Structure Prediction	65
4.8.3	Refinement of Vaccine 3D Structure	66
4.8.4	Validation of Vaccine 3D Structure	66
4.8.5	Proteasomal Cleavage Sites Prediction of MHC-ligands	68
4.9	Molecular Docking of Vaccine Protein with TLR-2 and its Structural Stability	68

4.10 Immune-simulation	70
4.11 Gene Cloning	72
4.11.1 Sequence Translation of Vaccine Protein and Codon Adap- tation	72
4.11.2 In-Silico Gene Cloning with the Expression System <i>E.coli</i> K12	73
5 Conclusion and Future Prospects	75
6 Bibliography	77

List of Figures

2.1	Arrangement of Spherical bacterial cells [2]	5
2.2	The phylogeny of the streptococcal species is obtained from a core set of 136 combined genes. Different color shades shows the major eight groups of species. The numbers on the branches indicates bootstrap support for each relationship [4].	7
2.3	By using electron microscope examining <i>S. gordonii</i> on a dentin slice of human. Premolars having single root were used to produce Human dentin slice. On sterile dentin slices <i>S. gordonii</i> was grown at a temperature of 37 °C for 6 h. Visualizing <i>S. gordonii</i> through electron microscope shows chains or pairs of clustered and spherical of <i>S. gordonii</i> [5].	9
2.4	Infectious diseases linked with <i>S. gordonii</i> [5].	10
2.5	Virulence factors associated with cell wall of <i>S. gordonii</i> [5].	11
2.6	Inflammatory responses mediated through <i>S. gordonii</i> lipoproteins [5].	12
2.7	Different types of vaccines against certain pathogens. Showing against which vaccine is licensed against which pathogens and when a certain type of vaccine was first developed [16].	16
2.8	Immune response generated to a vaccine [16].	17
2.9	RV process: (1) protein selection from genome sequence; (2) recombinant subunit vaccines production; (3) RV candidates evaluation: humoral and cellular responses of proteins and protection against lethal challenge [18].	18
3.1	Flowchart of methodology	33
4.1	Graphical Representation of Population Coverage by MHC restricted epitopes in World.	59
4.2	Graphical Representation of Population Coverage by MHC restricted epitopes in World.	59
4.3	Order of peptides in multivalent vaccine construct. Sequence of the epitopes and linkers was marked by different colors (Aqua colored residues: Cholera enterotoxin subunit B, yellow residues: B-cell epitopes; Pink residues: MHC-II restricted epitopes (HTL); Turquoise residues: MHC-I restricted epitopes (CTL), blue, green and orange residues: linkers.	61
4.4	Graphical Presentation of Vaccine I and Vaccine II	61
4.5	Secondary Structure of Vaccine I	65

4.6	Ramachandran Plot of Refined 3D Structure (GalaxyWeb) of Vaccine I	67
4.7	ERRAT Plot of 3D Structure (iTasser) of Vaccine I before refinement	67
4.8	ERRAT Plot of 3D Structure (GalaxyWeb) of Vaccine I after refinement	68
4.9	Protein-Protein interaction of Vaccine I (Ligand: Blue) with TLR2 (Receptor: Green) via ClusPro	69
4.10	In-Silico Immune Simulation of Vaccine with Adjuvant (Vaccine I) predicted through C-IMMSIM tool following Injection of Vaccine; (a) Antigen and immunoglobulins: antibodies production by antigen and their sub-division. (b) Cytokines: rise in the concentration of cytokines and interleukins. D shows the danger signal along with higher production of IL-2 (growth factor) (c) Total B-cell population count (d) Total B-cell population count per state (e) Total T-helper cell (CD4) count (f) Total population of TH cell per state (g) Total population count of T-cytotoxic cell count (h) Total TC cell population count per state.	72
4.11	In-silico Immune Simulation of Vaccine without Adjuvant (Vaccine II) predicted through C-IMMSIM Tool following Injection of Vaccine; (a) Antigen and immunoglobulins: antibodies production by antigen and their sub-division. (b) Cytokines: rise in the concentration of cytokines and interleukins. D shows the danger signal along with high production of IL-2 (growth factor)	72
4.12	Codon adaptation of Improved DNA.	73
4.13	In-Silico Gene Cloning of codon optimized vaccine with the Expression system E.coli K12; pET-28 (+). Vaccine clone was obtained by integrating the vaccine fragment into vector pET-28a (+) of the E.coli vector between the restriction sites BssSI (3665) and PciI (3224). Plasmid is shown in black color and DNA sequence of vaccine is shown in red color.	74

List of Tables

2.1	Scientific classification of <i>Streptococcus gordonii</i> [6]	8
4.1	List of Proteins based on their Sub-Cellular Localization	36
4.2	Uniprot ID, Virulence and Molecular Weight of Proteins	36
4.3	Antigenicity and Allergenicity of <i>S. gordonii</i> Surface Proteins	37
4.4	Physio-chemical Properties and Solubility of YSIRK_signal domain protein and Peptidoglycan D, D-transpeptidase FtsI	39
4.5	Function and Pathway Analysis of YSIRK_signal Domain Protein and Peptidoglycan D, D-transpeptidase FtsI	40
4.6	B cell epitopes of Selected Protein via ABCpred	41
4.7	Antigenicity, Allergenicity, Toxicity, and Sub-cellular Localization Evaluation of B-Cell Epitopes	42
4.8	MHC-I alleles binding peptides of YSIRK_signal domain-containing protein and Peptidoglycan D,D-transpeptidase FtsI computed via IEDB along with their antigenicity computed via VaxiJen 2.0	43
4.9	Antigenicity, Allergenicity, Toxicity, and Sub-cellular Localization of Selected MHC-I Restricted Epitopes	45
4.10	Physio-Chemical Properties of MHC-I Restricted Epitopes	46
4.11	Non-Digesting Enzymes of MHC-I Restricted Epitopes	47
4.12	MHC-II Alleles Binding Peptides of YSIRK_signal domain-containing Protein and Peptidoglycan D,D-transpeptidase FtsI computed via IEDB along with their antigenicity computed via VaxiJen 2.0	50
4.13	Antigenicity, Allergenicity, Toxicity, and Sub-cellular localization of selected MHC-I Restricted Epitopes	53
4.14	Physio-chemical properties of MHC-II Restricted Epitopes	54
4.15	Non-digesting enzymes of MHC-II Restricted Epitopes	54
4.16	Conservation analysis of epitopes of YSIRK_signal domain protein computed via IEDB epitope conservancy analysis tool	57
4.17	Conservation analysis of epitopes of Peptidoglycan D,D-transpeptidase FtsI computed via IEDB epitope conservancy analysis tool	58
4.18	Total Population Coverage of Designed Vaccine: Pakistan	58
4.19	Total Population Coverage of Designed Vaccine: World	59
4.20	Antigenicity and allergenicity of multivalent vaccine	62
4.21	Physio-chemical properties and solubility of multivalent vaccine	64
4.22	Predicted hydrogen bonding of docked vaccine I with TLR2 via PDBePISA	69
4.23	Predicted salt bridges of docked vaccine I with TLR2 via PDBePISA	69

Abbreviations

ABCpred	Artificial neural network based B-cell epitope prediction server
BLAST	Basic Local Alignment Search Tool
CAI	Codon Adaptation Index
DEG	Database of Essential Genes
DIANNA	DIAMinoacid Neural Network Application
EDGAR	Efficient Database framework for comparative Genome Analyses
ERRAT	Overall Quality Factor
GC	Guanine-cytosine Content
GRAVY	Grand Average of Hydrophaticity
HLA	Human Leukocyte Antigens
IC50	Half Maximal Inhibitory Concentration
IE	Infective Endocarditis
IEDB	Immune Epitope Database
i TASSER	Iterative Threading ASSEmbly Refinement
KEGG	Kyoto Encyclopedia of Genes and Genomes
MHC	Major Histocompatibility complex
MW	Molecular Weight
NA	Non-Allergen
NCBI	National Center for Biotechnology Information
NetChop	Neural network predictions for cleavage sites of the human proteasome
NT	Non-Toxin
PDB	Protein Data Bank

PDBePISA	Protein, Interfaces, Structures and Assemblies
PSIPRED	PSI-blast Based Secondary Structure PREDiction
RCSB PDB	Research Collaboratory for Structural Bioinformatics
RV	Reverse Vaccinology
<i>S. gordonii</i>	<i>Streptococcus gordonii</i>
TLR	Toll-like Receptors
TMHMM	Transmembrane Helices; Hidden Markov Model
trRosetta	Transform-restrained Rosetta
UniProt	Universal Protein Resource Using BLAST score Ratios
VFDB	Virulence Factor Database

Chapter 1

Introduction

Bacteria are small organisms with a single cell. They are found almost everywhere on Earth and are essential for the planet's ecological systems. Few species can survive under intense conditions of pressure and heat [1]. The human body is consists of numerous bacteria and it is actually estimated that it contains more bacterial cells than human cells. Many of them present in the body are not harmful, and few of them are beneficial. Only a small number of species cause disease [1].

Streptococcus belongs to the family Streptococcaceae, genus of gram-positive spherical bacteria or coccus within the order Lactobacillales (also called lactic acid bacteria), in the phylum Bacillota [2]. Bacteria belonging to this genus are found in almost all parts of the human body. Although streptococci dominate the mouth and are the main species found in the saliva and soft tissues of the oral cavity. In addition, streptococci have an essential role in the development of the microbiome of mouth [3]. These bacteria are the ones that are found in the mouths of newborns and are treated as the primary colonizer, allowing other types of bacteria to assemble [3].

Many species of streptococci live in humans without causing symptoms; however, some types can lead to a variety of serious illnesses - from sinus infection to pneumonia. Particularly, infections caused by viridans streptococci worsen when bacteria invade other parts of the body. When bacteria go into the bloodstream, it results in serious infection in the heart's lining, called infective endocarditis.

These heart infections can be lethal and often require clinical care and antibiotic treatment. Sufferers with weak immunity or already having problems of cardiac valve are at greater risk [3]. Streptococcus strains are the initial colonizers that inhabit the oral cavity and can be found after birth and therefore, play significant role in the accumulation of microbiota of mouth.

Oral streptococci induce collection of adherent molecules which allow them to properly collect distinct tissues in the oral cavity. Also, they have an amazing ability to digest carbohydrates by fermenting thus producing acids as a by-product [4]. Development of dental caries is caused by aciduric species like *Streptococcus mutans* by excessive acidification of the oral cavity [4]. Whereas, less acid-tolerant species including *Streptococcus salivarius* and *Streptococcus gordonii* produce a significant amount of alkaline expression and play a significant role in maintaining acid-base physiology of the mouth [4].

Another significant feature of oral streptococci is their capability to produce hydrogen peroxide that retards the growth of *S. mutans* [4]. Therefore, oral streptococci can be advantageous to the host by means of generating molecules that retard the growth of pathogenic species. When the mutualistic and pathogenic streptococci residing in the mouth can ultimately reach the human blood system, they cause systemic illness and infections such as endocarditis [4].

S. gordonii, a Gram-positive and mutualistic bacterium normally found on the skin, mouth, and intestines. It is an opportunistic pathogen and cause diseases including infective endocarditis and apical periodontitis [5]. *S. gordonii*, an initial colonizer, adhere itself easily to adhesive tissues of host that includes both the dental surface and cardiac valves, forming plaque called biofilms. *S. gordonii* enters the inside of tooth i.e., root canal and vessels of blood, after which they interact with numerous immune (leukocytes) and non-immune cells of the body. *S. gordonii* cell wall consists of lipoproteins, LTAs (lipoteichoic acids), peptidoglycans, repetitive serine-rich adhesion molecules, and cell wall proteins that were characterized by host receptors [5]. All these proteins are associated with virulence and immune regulatory actions that induce inflammation. So, the elements of cell wall of *S. gordonii* act as virulence factors which progressively develop diseases

with a strong participatory response [5]. *S. gordonii* is incorporated into some of the earliest colonizers of periodontal origin. *S. gordonii* shows a high affinity for molecules in the salivary pellicle (or adhesive) in the dental areas [6]. Therefore *S. gordonii* can quickly colonize areas of clean teeth, and *S. gordonii* and their related organisms account for a high percentage of bacterial biofilm, up to 70%, that forms in clean dental areas. It is usually harmless in the mouth but when it gains access systemically (blood), *S. gordonii* can cause endocarditis. *S. gordonii* also formed a substratum that attaches to later colonizers in the dental area and could alter the pathogenicity of these second colonizers by using interspecies communication methods [6].

Oral streptococci comprising *S. gordonii* (Sg) and *S. sanguinis* (Ss), are one of the utmost common colonies of biofilms in dental areas, called as dental plaque. Sg and Ss can attach to parts of the salivary pellicle and oral bacteria by using a wide range of adhesion proteins that are expressed in the area of cell [7]. This interaction is thought to help initiate and further the formation of dental plaque. Additionally, Sg and Ss can invade the bloodstream and are main cause of occasional, yet deadly bacterial infectious disease called infective endocarditis (IE).

These are α -hemolytic oral streptococci and have recently been determine in neuropaenic infections of blood [7]. Infectious endocarditis is a not common but serious illness that is observed worldwide. The majority of the pathogenic specie seen in endocarditis is from the oral cavity and may arise from daily dental procedures or invasive procedures [8]. Infective endocarditis maybe caused by *S. gordonii* after the dental abscess drain out [8].

1.1 Aim and Objectives

Streptococcus gordonii are oral bacteria that initiate dental plaque formation and cause infective endocarditis. It is an emerging pathogen and causes infections. No vaccine developed to control this pathogen. So, the aim of study was to identify potential vaccine candidates and design vaccine models against *Streptococcus*

gordonii by using a reverse vaccinology approach from the analysis of publicly available data of 13 complete strains.

1. To identify core genomes of all the strains of *Streptococcus gordonii*.
2. To identify potential vaccine candidates.
3. To design a vaccine against *Streptococcus gordonii* by using reverse vaccinology.

Chapter 2

Literature Review

Bacteria are very small living cells, and some are considered to be the smallest possible size of an organism that reproduces independently. The common forms are circles, rods, curved or curved rods, and spirals [2]. Oval or round shaped bacteria are called cocci. Spirilla are the spiral-shaped bacteria. Among cocci, diplococci (pairs), streptococci (chains), and staphylococci (irregular clusters of bacteria) are found [2].

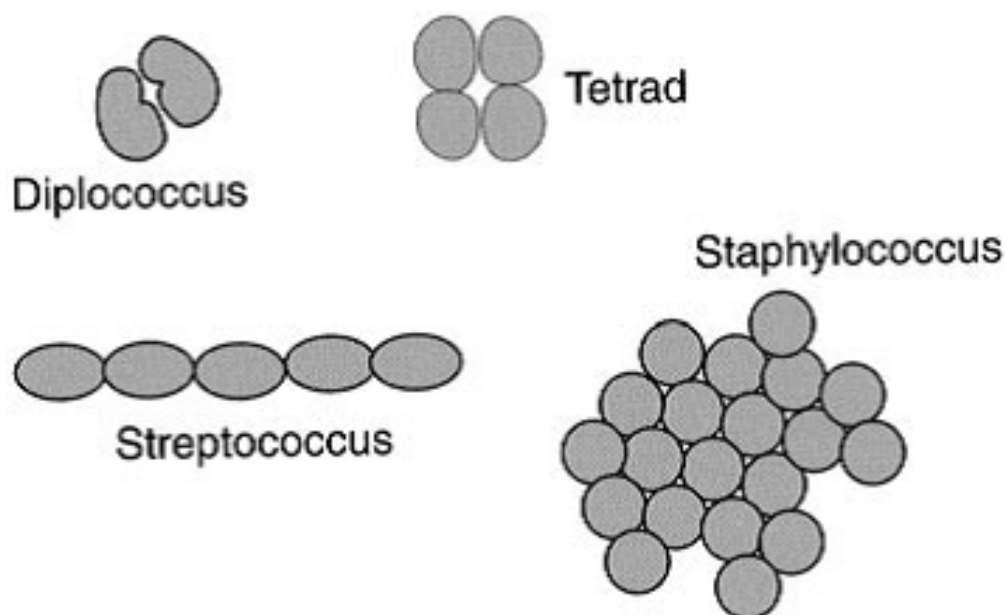


FIGURE 2.1: Arrangement of Spherical bacterial cells [2]

The mouth provides an ideal habitat for colonizing fungi, protozoa, archaea, and bacteria because of its warmer, wet, and nutrients rich surroundings. The organism

usually colonizes in the form of a complex structure called dental plaque or biofilm [4].

2.1 Streptococcus

Bacteria belonging to the genus streptococcus are Gram-positive that are commonly organized in the form of chains with oval cells connecting end to end. Besides harmless species, the genus also comprises important pathogens of humans [2]. Medically beneficial streptococci lack characteristics like acid-fast and spore formation and are non-motile [2]. Streptococcus genus is organized as Gram-positive bacteria, that may be cocci or spherical-shape, and in the form of clustered pairs or chains [5]. Depending on hemolysis on blood agar plates Streptococci are divided into three groups (Brown's classification): α -hemolysis (incomplete hemolysis), β -hemolysis (complete hemolysis) and γ -hemolysis (no hemolysis) [4]. Phylogenetic approaches further sub-divided streptococcus into group of eight consisting of bovis, anginosus, salivarius, pyogenic, sanguinis, downei, mutans, and mitis [4].

2.2 Oral Streptococci

Our mouth is home to hundreds of bacterial species. These bacteria play important role in digesting food, preserving the immune system, and preventing infections [3]. But some species of bacteria such as oral streptococci also contribute to the development of plaque and can sometimes lead to infection [3].

2.2.1 Taxonomy of Oral Streptococci

Streptococci are present in almost every part of the human body and are the dominant species found in the oral cavity and upper respiratory tract of humans. Genus Streptococcus were initially classified into 3 groups depends on hemolysis on blood agar plates including complete (β), partial (α) and non-hemolytic (β) [4]. Oral

streptococci were named *Streptococcus viridans* because when they were cultured on blood agar plates, they show partial hemolysis resulting in a green coloration of colonies [4]. By using a phylogenetic approach, classification of the streptococci into eight distinct groups includes sanguinis, mitis, salivarius, anginosus, mutans, downei, bovis and pyogenic [21].

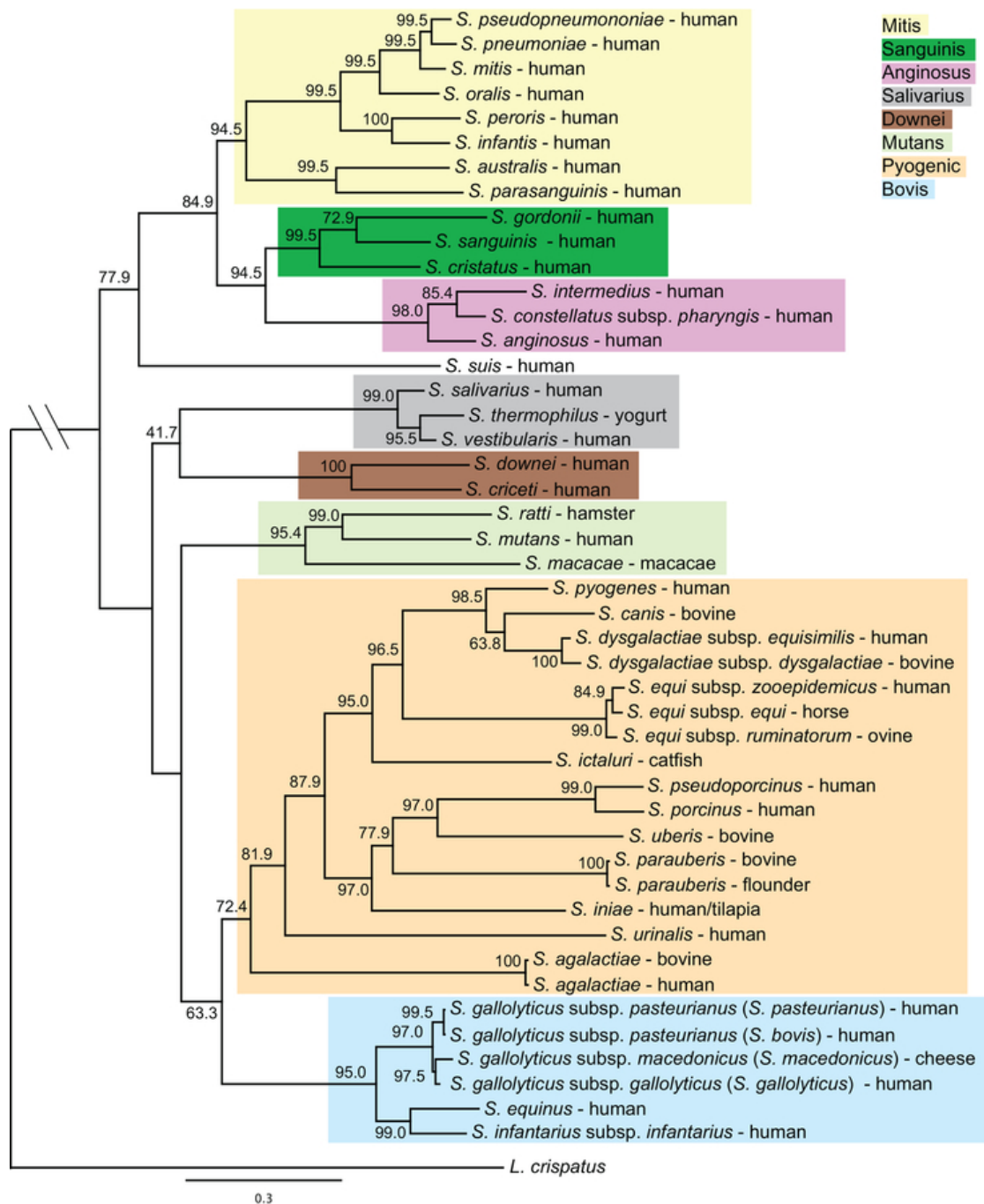


FIGURE 2.2: The phylogeny of the streptococcal species is obtained from a core set of 136 combined genes. Different color shades shows the major eight groups of species. The numbers on the branches indicates bootstrap support for each relationship [4].

2.3 Viridans Streptococci

Viridans streptococci are α -hemolytic and plays essential role in the formation of dental plaque. The species of viridans streptococci are the first colonizers of the teeth, mainly *Streptococcus gordonii*, *Streptococcus sanguinis*, and *Streptococcus mitis* [2] [3]. Viridans streptococci comprise species of the common flora of the human buccal cavity. They cause several diseases including purulent infections, endocarditis, meningitis, and septicemia [9]. Species of the common flora of the oral and nasopharyngeal cavities have basic bacteriologic characteristics of streptococci but lack certain antigens, toxins, and virulence of other groups. Although the viridans group includes a wide variety of species they are usually not identified in clinical practice because there is little difference in their medical significance [3]. Their diagnosis is based on a variety of physiological and biochemical characteristics. The identification of species of viridans streptococci is helpful in cases of infectious endocarditis and positive blood cultures cases and to assess the involvement of a particular strain in infection [9].

2.4 *Streptococcus gordonii*

S. gordonii is a mutualistic, non-pathogenic bacterium found in humans, involving the buccal cavity, upper respiratory tract, skin, intestines [5]. In humans and animals, it mainly inhabits mucosal and dermal surfaces including the mouth as well as lives in soil, water, food and plants [10].

TABLE 2.1: Scientific classification of *Streptococcus gordonii* [6]

Level	Taxonomy
Domain	Bacteria
Phylum	Bacillota
Class	Bacilli
Order	Lactobacillates
Family	streptococcaceae
Genus	Streptococcus

Specie *S. gordonii*

S. gordonii, is member of the group of viridians bacteria (α -hemolytic-sanguinis), which live mainly in the buccal cavity of animals and human. But, it is an opportunistic pathogen too which is also involved in causing various diseases and infections [5]. *S. gordonii*, the first colonizer on the surface of the tooth, can combine with numerous other microorganisms of buccal cavity, and contribute to the progression of periodontal infectious disease and tooth decay [11].



FIGURE 2.3: By using electron microscope examining *S. gordonii* on a dentin slice of human. Premolars having single root were used to produce Human dentin slice. On sterile dentin slices *S. gordonii* was grown at a temperature of 37 °C for 6 h. Visualizing *S. gordonii* through electron microscope shows chains or pairs of clustered and spherical of *S. gordonii* [5].

Recently, next-generation metagenomic sequence studies manifest, *S. gordonii* is present in sufferer with apical periodontitis or dental decay and also in cardiac valves in sufferer with infectious endocarditis [5]. By entering the bloodstream through oral bleeding, *S. gordonii* can lead to endocarditis [8]. *S. gordonii* also attaches to the surface of many host cells, causing the onset of disease through inflammation [5].

2.4.1 Pathogenicity

S. gordonii, commensal bacteria is present in space of the oral cavity and skin of the human body. It is mutualistic as well as emerging pathogen related with many infections and other diseases.

S. gordonii also enter the blood through dental trauma and tooth decay and spread into various body organs thus give rise to systemic infections and diseases comprising empyema, infective endocarditis, pyogenic spondylitis, perihepatic abscesses [5].

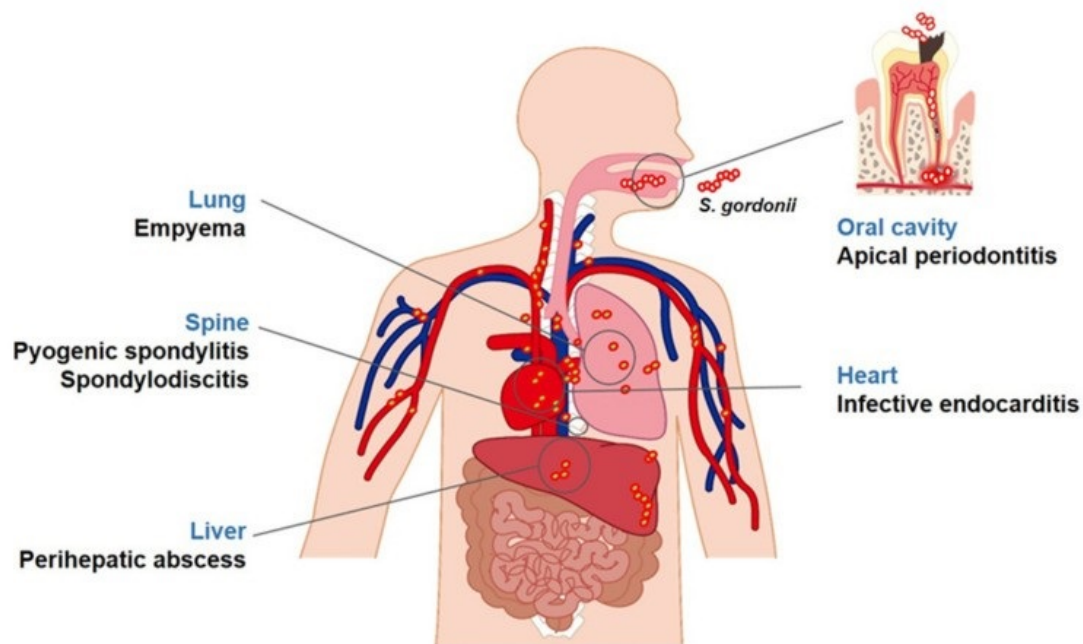


FIGURE 2.4: Infectious diseases linked with *S. gordonii* [5].

Bacteria cell wall play essential role in its survival and growth [5]. The Gram-positive bacteria has thick peptidoglycan (PGN) and many other elements comprising wall teichoic acid (WTA), lipoteichoic acid (LTA), lipoproteins, and cell-anchoring glycoprotein in its cell wall that are characterized by host cell receptors.

They are involved in virulence and immune regulatory processes that induce inflammation. Dimeric receptors containing TLR2 and TLRx recognize the lipoproteins and LTA. SRR adhesins are important for the binding of *S. gordonii* to host cells through sialylated glycans. Nucleotide oligomerization domain (NOD), an intracellular receptor, recognized peptidoglycans. Thus, cell wall elements of

S.gordonii act as virulence factors that progressively develop diseases with a strong participatory response [5].

The major virulence factors like lipoprotein of *S.gordonii*, are directly recognized by heterodimers, composed of toll-like receptors TLR2 along with TLR1 or TLR6 on host cell including dental pulp cells, dendritic cells, valve interstitial cells and macrophages. Following activation of TLR2, an adaptor molecule of TLR2, myeloid differentiation primary response 88 (MyD88), mediates the activation of transcription factor that is nuclear factor-kappa B (NF- κ B), results in the production of pro-inflammatory cytokines and chemokines, maturation of cell, and infiltration of immune cells into lesions. These processes are involved in inducing inflammatory responses and thus results in development of diseases like apical periodontitis or infective endocarditis. [5]

Proteins expressed by *S. gordonii* cell wall includes collagen-binding domain protein (CbdA), SspB, Streptococcal surface protein (Ssp) A, and serine-rich repeat (SRR) glycoprotein, including Hs antigen (Hsa) and gordonii surface protein B (GspB) [5].

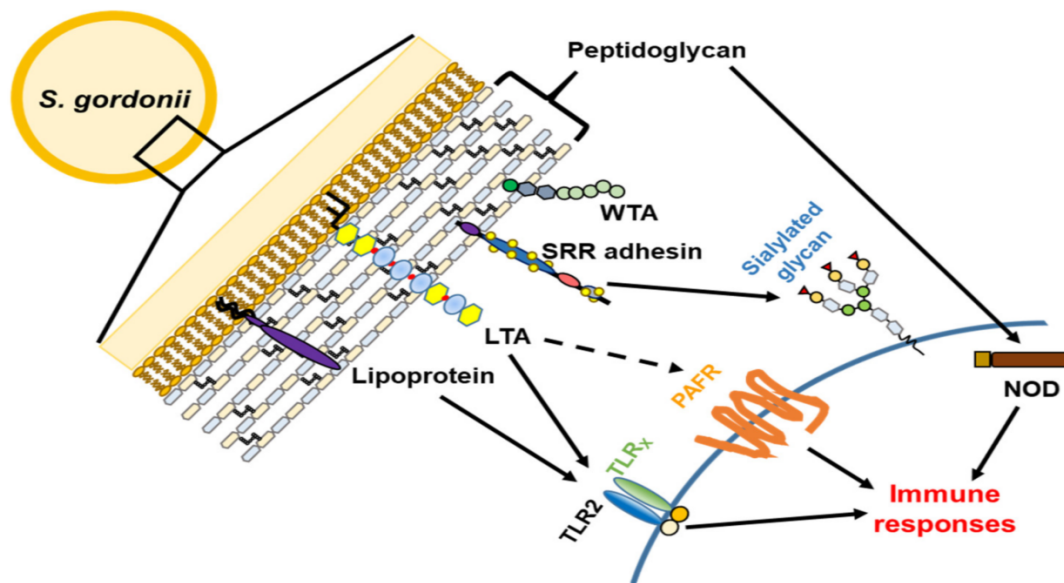


FIGURE 2.5: Virulence factors associated with cell wall of *S. gordonii* [5].

S. gordonii cell wall proteins freely attaches to thrombocytes, red blood cells, dendritic cells (DCs), and monocytes can cause acute immune response in humans.

So, knowledge of the interaction of *S. gordonii* cell wall components with host cells is necessary to describe its overall pathogenesis and also for the application of treatment methods and to prevent *S. gordonii*-mediated infectious diseases [5].

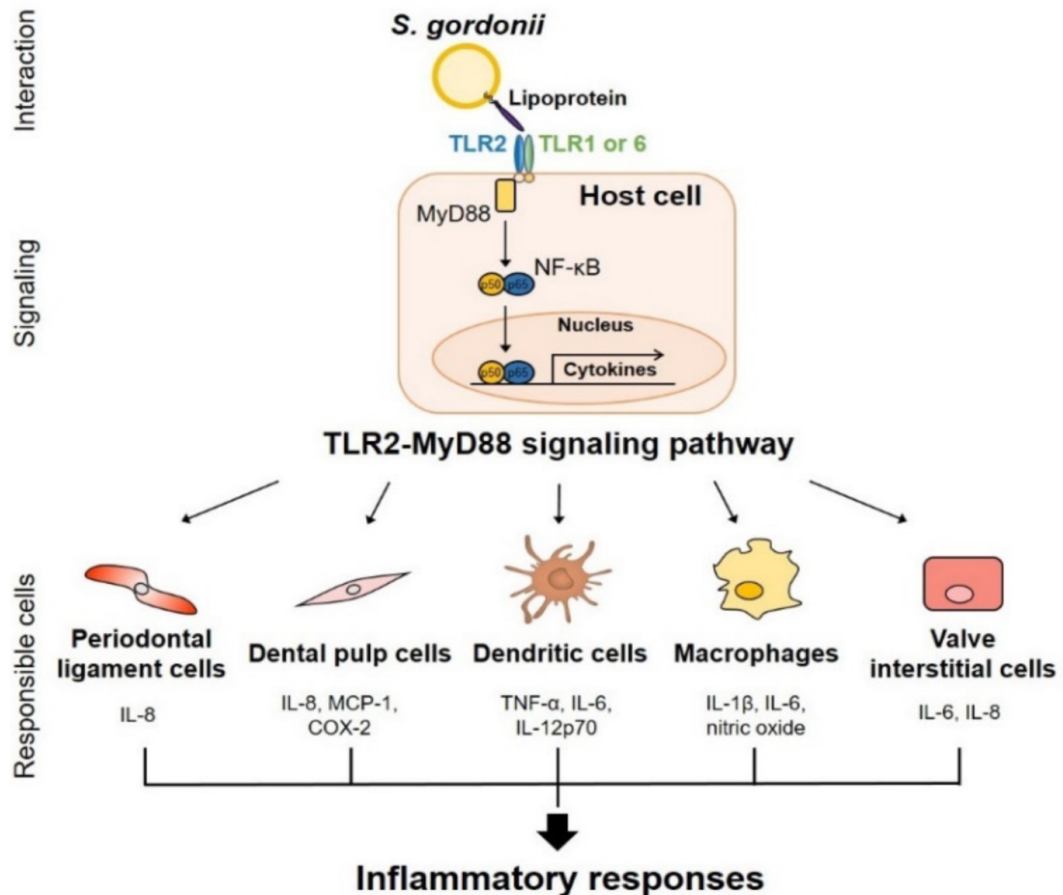


FIGURE 2.6: Inflammatory responses mediated through *S. gordonii* lipoproteins [5].

2.5 Infective Endocarditis

The inflammation of the endothelium of the heart is known as infective endocarditis (IE) [12]. IE is an orphan but serious disease affecting low- and high-income countries alike [8]. Annual cases of 3-10 / 100,000 with a death rate of up to 30% in 30 days occur due to infective endocarditis [12].

Endocarditis is endocardium infection. The endocardium is the inner lining of the heart chambers and heart valves. *S. aureus* is the most common pathogen that causes IE in most researches in 26.6% of overall cases, whereas viridans group

streptococci cases are 18.7%, other streptococci cause 17.5%, and enterococci cases are 10.5% [12]. Streptococcus, Staphylococcus, and Enterococcus species together account for about 79-90% of overall cases of IE around the world [8].

Infective endocarditis (IE) remains low but highly contagious disease [12]. A category of viridans streptococci (VGS) bacteria have low pathogenicity and are common in the buccal cavity, gastrointestinal tract, respiratory tract, and female genitals [8]. This group is further divided into six major subgroups: *S. mitis*, *S. mutans*, *S. salivarius*, *S. anginosus*, *S. bovis*, and *S. sanguinis* [4]. Group *S. sanguinis* is contagious. *S. sanguinis* including *S. parasanguinis*, and *S. gordonii* cause infectious diseases including IE [8].

S. gordonii has been rarely reported as the cause of IE in researches. *S. gordonii* plays a significant role in the alkalization of the buccal cavity with the generation of a protective biofilm [8]. Whereas, in the bloodstream of human *S. gordonii* appear to possess pathogenic virulency for IE development [8].

2.5.1 Prevalence of Infective Endocarditis in Pakistan

In Pakistani, according to clinical observation, location of the valvular heart infection, and types of pathogenic microorganisms were common with infectious endocarditis [15]. However, compared with patients with infectious endocarditis in economically developed countries, the significant variation in population being studied included younger age at baseline, higher incidence among women, and a higher number of cases involving congenital heart disease [15].

The appearance of arrhythmias and neurological complications, and cardiac heart failure cause mortality in persons suffering with infective endocarditis [14]. Patients with a history of heart surgery and those with neurologic and embolic disorders had a low prognosis [15]. The presence of kidney failure and heart failure also predicted the risk of death. Existing diagnostic methods for endocarditis can be used in Pakistani patients, which allows for accurate diagnosis of infected patients even in cultures of negative blood caused by regular self-administration of antimicrobial agents before introduction [15].

2.5.2 Antibiotic Treatment

Infectious endocarditis was contagious before the arrival of antibiotics [12]. There is a growing experience of using ceftriaxone as synergistic enterococcal endocarditis; as a result, amoxicillin, and ceftriaxone are recommended in European guidelines and are especially useful in patients with renal impairment or with severe bleeding are best taken after consultation with a pathologist [12]. Where possible, it is advisable to keep an intermediate line in place due to the long duration of antibiotic treatment [12].

2.5.3 Challenges in IE

Infectious endocarditis is defined as an inflammation within the heart and is a deadly disease in the area of cardiovascular diseases [13]. The challenges of infectious endocarditis are important. It differs in pathogenecity, clinical symptoms, and treatment. *S. aureus* leads to a more severe form of the illness, usually in the individuals at risk or the aged persons [13]. There is no finance and facilities for research, and lack of randomized controlled clinical trials that guides performance. Long-term conflicts such as the role of antibiotic prophylaxis or the duration of surgery have not yet been resolved [13].

2.6 Vaccinology

Vaccination is the basis of public health policy and appears to be less expensive when used to protect children's health from the pathogen. A vaccine is a biological product that can be used to safely initiate an immune reaction and provide defense against infectious disease upon exposure to a pathogen [16]. To attain this, the vaccine should possess antigens that are obtained from pathogen or synthetic to show pathogen elements. An important component of many vaccines to induce immune response is one or more protein antigens that provide defense against pathogen [16]. However, polysaccharide antigens may be contribute to the immune response and provide guide for vaccines development to stop many

bacterial infections to infect the body, such as meningitis and pneumonia caused by *S. pneumoniae*, since the year 1980s [16].

2.6.1 Types of Vaccine

Vaccines can be normally organized as live or non-live to separate those vaccines which contain subspecies of pathogen-related to the one which composed of only parts of the pathogenic organisms or killed organisms.

Besides living and non-living, 'traditional' vaccines, and many other forms have been produced in the last few year, comprising nucleic acid-based RNA, viral vectors, viral-like particles and DNA vaccines [16].

2.7 Immune Response

Humoral immunity is mediated by B cells through antibodies production cellular immunity is mediated by T cells i.e., adaptive immune response. All vaccines that are used, provides protection by the production of antibodies [16]. Different types of antibodies are important in induction of protection through vaccine. This is proved through three main areas: immunodeficiency states, passive protection studies, and immunological data [16].

In a muscle, vaccine is injected. When it is injected, dendritic cells are activated by pattern recognition receptors (PRRs) and protein antigen is taken up by dendritic cells and enters the lymph node. MHC molecules present on dendritic cell activates T cells through their T cell receptor (TCR). T cells cause B cell development through the B cell receptor in the lymph node. This causes maturation of the antibody reaction so to enhance antibody efficiency and produce different isotypes of antibody [16].

Plasma cells that are short lived, produce antibodies specific to the vaccine protein and within the next 2 weeks resulting in a rise in serum antibody levels. Memory B cells are also secreted that mediate immunity [16]. Plasma cells that are long







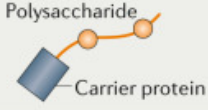
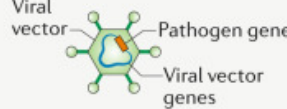
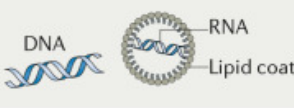

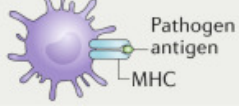
Type of vaccine		Licensed vaccines using this technology	First introduced
Live attenuated (weakened or inactivated)		Measles, mumps, rubella, yellow fever, influenza, oral polio, typhoid, Japanese encephalitis, rotavirus, BCG, varicella zoster	1798 (smallpox)
Killed whole organism		Whole-cell pertussis, polio, influenza, Japanese encephalitis, hepatitis A, rabies	1896 (typhoid)
Toxoid		Diphtheria, tetanus	1923 (diphtheria)
Subunit (purified protein, recombinant protein, polysaccharide, peptide)		Pertussis, influenza, hepatitis B, meningococcal, pneumococcal, typhoid, hepatitis A	1970 (anthrax)
Virus-like particle		Human papillomavirus	1986 (hepatitis B)
Outer membrane vesicle		Group B meningococcal	1987 (group B meningococcal)
Protein-polysaccharide conjugate		<i>Haemophilus influenzae</i> type B, pneumococcal, meningococcal, typhoid	1987 (<i>H. influenzae</i> type b)
Viral vectored		Ebola	2019 (Ebola)
Nucleic acid vaccine		SARS-CoV-2	2020 (SARS-CoV-2)
Bacterial vectored		Experimental	–
Antigen-presenting cell		Experimental	–

FIGURE 2.7: Different types of vaccines against certain pathogens. Showing against which vaccine is licensed against which pathogens and when a certain type of vaccine was first developed [16].

lived continuously produce antibodies for a long time period and are present in the bone marrow as well as in blood. CD8⁺ memory T cells suddenly proliferate when they contact a pathogen, and CD8⁺ effector T cells cause the removal cells that are infected [16].

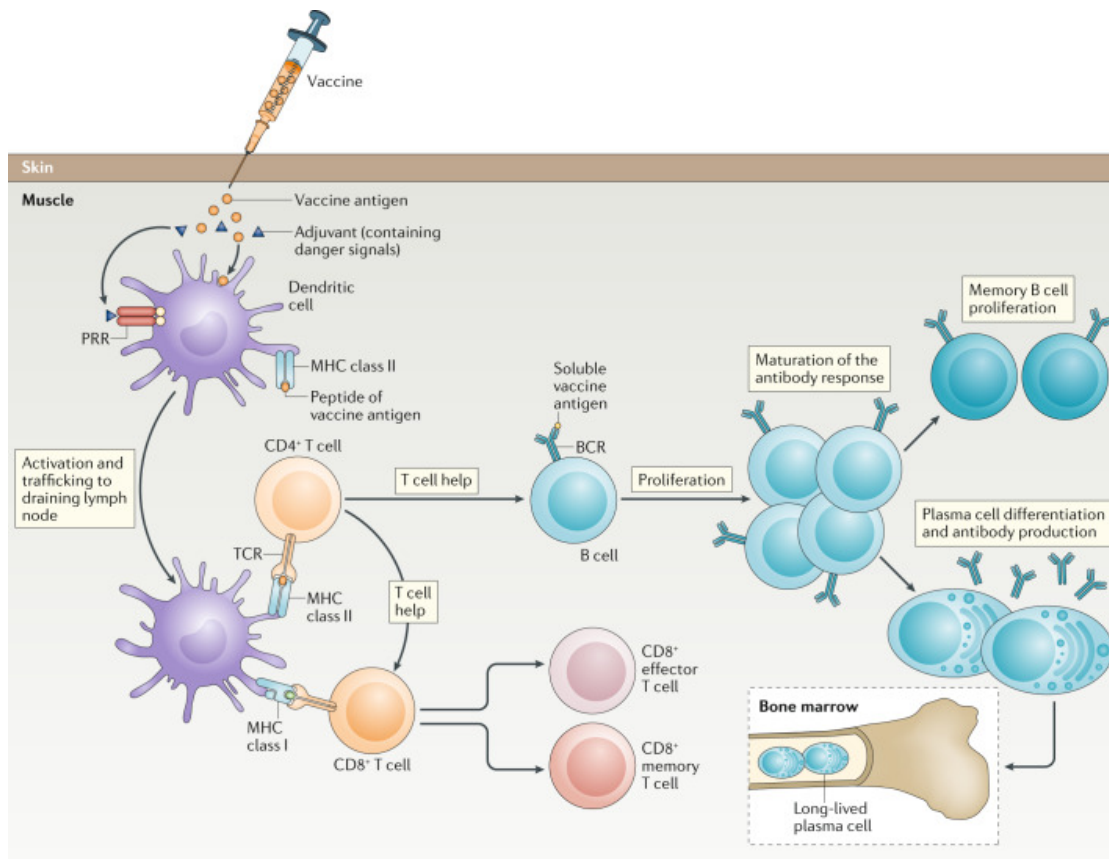


FIGURE 2.8: Immune response generated to a vaccine [16].

2.8 Reverse Vaccinology

One of the genome-based approaches is known as reverse vaccinology (RV). Reverse vaccinology depends on the use of computational methods and tools to identify candidates for further experimentation/testing of the vaccine, refinement of which is essential for its full use as opposed [17].

In the 1990s, reverse vaccinology was first developed by Rappuolito identify potential antigens for a vaccine against the B strains of *Neisseria meningitidis* (Group B meningococcus – MenB) [17]. Through the RV approach, the vaccine is developed by acquiring computer based identification of protein sequences from the pathogenic organism and then selecting candidate antigens that is known as potential vaccine candidates (PVCs) [19].

RV provides two major advantages as compared to the development of traditional vaccine: Candidate antigens are being identified from pathogen growth, and an antigen is identified for testing of vaccine of its purified quantity [19].

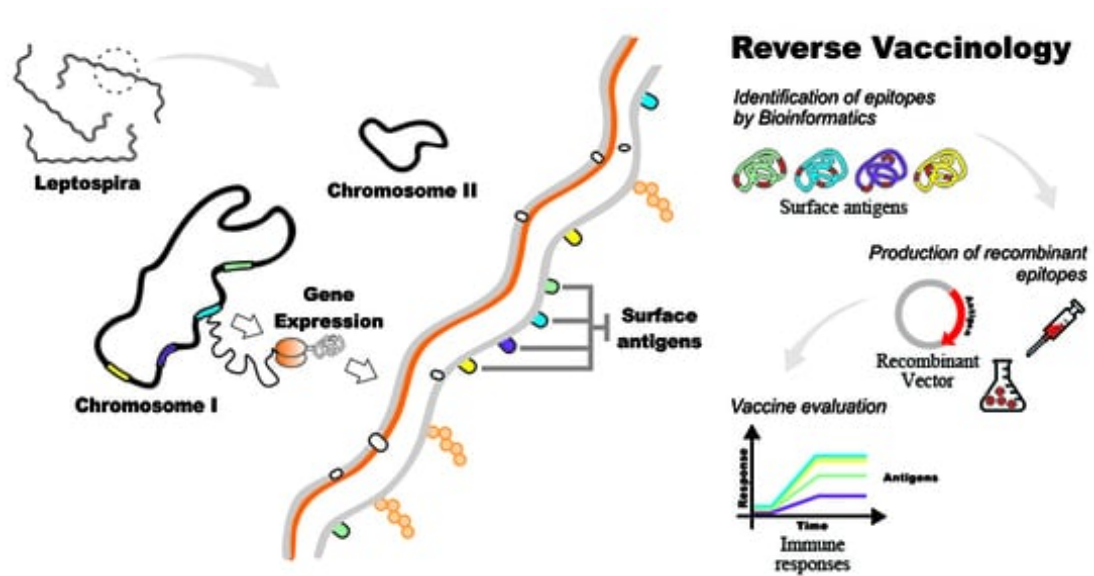


FIGURE 2.9: RV process: (1) protein selection from genome sequence; (2) recombinant subunit vaccines production; (3) RV candidates evaluation: humoral and cellular responses of proteins and protection against lethal challenge [18].

2.9 Pan-Genome

The analysis of entire gene set present in the given data set is pan genome analysis. In bacterial taxonomy; pan-genome is one out of several new innovative tools. By using pan-genomic analysis, species can be defined again, and new species can be described.

By applying genomic and pan-genomic analyses bacterial species can be reclassifying. This will be beneficial in the medical microbiology future [20]. It includes core genomes, accessory genomes and specific or unique genes. It identifies the genomic diversity present between the strains of distinct species [20].

2.10 Scope of Study

Vaccine is used to prevent the disease caused by microorganism. There is no vaccine against *S. gordonii*. So, it is helpful in pharmaceutical industries to develop vaccine against *S. gordonii* by using reverse vaccinology (RV) approach. It will also helpful to raise the economy of the country. Antibiotic treatment is available against infection caused by *S. gordonii*. But antibiotic treatment is not long lasting

and infection can be occurred again. Vaccine has long lasting impact on the organism to prevent infectious diseases. It is a preventive way to stop the onset of disease. So by developing vaccine, one can develop immunity against *S. gordonii* and prevent infections and diseases like infective endocarditis (IE).

2.11 Impact on Society

S. gordonii is a mutualistic, non-pathogenic bacterium found mostly in mouth, skin of humans. It is emerging pathogen and cause serious infections. When *S. gordonii* enter the blood through dental trauma and tooth decay and spread into various body parts, give rise to systemic infections and diseases comprising empyema, apical periodontitis, infective endocarditis, pyogenic spondylitis, perihepatic abscesses [5]. People with low immunity are at high risk of infection. This will increase the symptoms and suffering of patients. So the vaccine develop against *S. gordonii* ensures to increase the immunity against *S. gordonii*. This would be helpful to prevent the burden of disease and reduce the suffering of person after exposure to disease.

2.12 Gap Analysis

S. gordonii is the emerging pathogen and causes serious infectious diseases. These infections were treated with antibiotics. Antibiotic treatment is not the permanent cure. Infection can occur again and again because of low immunity. Vaccine increases the immunity of a person against certain diseases. There is no vaccine to immune the patients permanently against the opportunistic pathogen *S. gordonii*.

2.13 Research Question

1. How many genetic variants of *S. gordonii* were being identified and what is the common gene that is present in all the variants?

2. What makes it opportunistic pathogen and cause diseases?
3. What are the essential proteins for the survival of *S. gordonii* and could be targeted to develop vaccine against *S. gordonii*?
4. How the vaccine is developed against *S. gordonii* by reverse vaccinology?

Chapter 3

Materials and Methods

As discussed earlier, there is no vaccine for *S.gordonii*. Therefore reverse vaccinology approach would be used to identify candidates (antigens) for vaccine development using computational tools and methods.

3.1 Vaccine Target Prediction

3.1.1 Sequence Retrieval and Genome Selection

All the sequences of genes and proteins were obtained from [NCBI](#). This will predict the total strains of the organism including complete, contig and scaffold [24]. Only strains with complete genome were selected. Genome Assembly and Annotation report shows total 91 strains of *S. gordonii*. Out of 91 strains of *S.gordonii* 13 are complete, 65 are contig, and 13 are scaffold. The thirteen strains (complete) of *S. gordonii* are included in this study that has complete genome.

3.1.2 Identification of Core Genome

The core genome of *S. gordonii* was identified using [EDGAR 2.3 software](#) by selecting reference strain which is “Challis substr. CH1” on the basis of its release date

that was compared with other 13 strains. The core genomes that were common in all the strains were selected.

3.1.3 Homologous and Non-homologous Protein Identification by using Subtractive Genomic Analysis

Sequence between the host and pathogen proteome are subtracted to identify proteins important for pathogen but not present in host by using subtractive genomics. FASTA sequence of all proteins was obtained from NCBI and then they were BLAST against humans to identify homologous and non-homologous proteins.

3.1.4 Identification of Essential genes

These non-host homologous proteins were inserted in [DEG database](#): database of essential genes to find out the essential proteins by using thresholds. The default parameters will be selected, E value = 0.0001, bit score > 100, and identity > 25%.

3.2 Collection of Surface Proteins

The subcellular localization of the proteins was identified using [CELLO](#) tool. Based on the sub-cellular localization of proteins in *S. gordonii*, only surface proteins were selected [22]. Five surface proteins were present in *S. gordonii* and were selected for designing vaccine against *S. gordonii*.

These include YSIRK signal domain protein, Peptidase C51 domain-containing protein, AraC family transcriptional regulator, Glycosyl transferase, Peptidoglycan D, D-transpeptidase FtsI, AraC family transcriptional regulator. Out of these five proteins, first two proteins were extracellular and other three proteins were membrane proteins.

3.2.1 Sequence Retrieval of Proteins

After the collection of proteins, first step is the sequence retrieval of proteins sequence using [Uniprot](#) [22]. Virulence of the protein targets was identified using VFDB Molecular weight of proteins was calculated from [ProtParam tool](#).

3.2.2 Screening Antigenicity and Allergenicity of *S. gordonii* Proteins and Protein Selection

To predict the antigenicity of proteins, all the five proteins were filtered by using the [VaxiJen v2.0 webserver](#) at threshold 0.4. After submitting the FASTA sequences of all five structural proteins to VaxiJen v2.0 server, the antigenicity of proteins will be predicted depending on the physiochemical features of the protein. This analysis led to the selection of protein whose antigenicity is above 0.4.

Effective prediction of 70-89% was provided by this server [29]. The allergenicity of the proteins was predicted through [AllerTOP v.2.0](#). This is bioinformatics server for allergenicity prediction. The proteins that were non-allergen were selected for further analysis.

3.2.3 Structure Analysis by Screening of Physio-chemical Properties of Selected Proteins

The selected proteins that were antigenic and non-allergen were subjected to structure analysis for determining the physiochemical properties. To predict the physiochemical properties of the proteins, the tool use was [ProtParam](#).

This tool predicts the molecular weight of protein, number of amino acid, aliphatic index of protein, pI value, instability index of protein, aliphatic index of protein, and estimation of half life of the protein. The solubility of the protein was predicted through [SoluProt](#). To predict disulphide bonds in both proteins [DIANNA 1.1](#) was used.

3.2.4 Function and Pathway Analysis of Selected Protein

The function and pathway analysis of selected proteins was done through [uniprot](#) by using uniprot ID of the protein. The pathway of the protein was also predicted using [KEGG](#) web tool.

3.3 Epitope Prediction

One of the important steps in immunoinformatics study is the epitope selection. Multi-specific and broad-based epitopes should be selected. Multi-specific epitopes are the epitopes that can be identified from many proteins originating from a single pathogen and broad-based epitopes are a series of epitopes that are derived from a single protein.

Factors that influence epitope selection include the ability of the epitope to attach to the appropriate MHC molecule, cellular presentation ability of epitope, and the repertoire of T cells should possess the ability to differentiate between MHC-epitope complexes [26]. Epitopes prediction was done through [ABCpred](#) and [IEDB](#) analysis resource.

3.4 Prediction of B-Cell Epitopes (LBL) via Immunoinformatics

B-lymphocytes are responsible for secreting antibodies that provides long term immunity to body against diseases and infections [22]. B lymphocytes plays important role in adaptive immunity so, they are major part of the immune system.

B-lymphocytes (LBL) epitopes were predicted through online webserver [ABCpred](#). The fasta sequences of proteins were entered in the ABCpred at threshold 0.5 and length of epitope is 16. In this way, the B-cell epitopes of the antigenic protein were determined as major part of the immune system [24].

3.4.1 Antigenicity, Allergenicity, Toxicity, Mol. Weight and Sub-Cellular Localization Profiling of Epitopes

After the prediction of B-cell epitopes, [VaxiJen v2.0](#) tool at threshold of 0.4 and [AllerTOP v2.0](#) tool was used for the prediction of the antigenicity and allergenicity of the epitopes [24]. The epitopes which were antigenic and non-allergen were selected. Molecular weight of the epitopes was predicted through [ProtParam](#) tool. The toxicity of the epitopes was predicted through [ToxinPred](#) tool. The sub-cellular localization of epitopes was also predicted through [TMHMM-2.0 web server](#).

3.5 Prediction of T-cell Epitope via Immunoinformatics

T-Cell epitope prediction tool of [IEDB](#) analysis resource was used to evaluate T-cell epitopes. This tool predicted the peptides that were bind to MHC-I molecules and MHC-II. This tool determines the binding affinity of peptides towards the MHC-I and MHC-II on the basis of their IC50 value.

Peptide having IC50 value less than 50 nM, possess high affinity of binding towards MHC-I and MHC-II, peptide having IC50 between 50-500nM shows midrange affinity and peptides having IC50 between 500-5000nM indicates lowest binding affinity [22]. IC50 score is inversely related to percentile rank. On the basis of IC50 score, MHC-I and MHC-II epitopes were selected.

3.5.1 MHC-I Binding Epitopes (Cytotoxic T-lymphocytes - CTL)

The MHC-I binding epitopes (CTL) for two selected proteins were evaluated by using [IEDB](#) tool. Protein fasta sequence was entered in this MHC-I binding predictions tool. The prediction method selected was ANN 4.0 and the MHC resource

specie was human. All the HLA-A* alleles were selected and the predicted length of epitope was 9-mer [23].

3.5.2 MHC-II Restricted epitopes (Helper T-cell Lymphocytes (HTL) Epitopes)

Helper T-cells plays an essential role in all adaptive immune responses. HTL stimulates B-cells to produce antibodies and help macrophages to engulf and absorbs the pathogens. In addition to this, HTL also activates the CTL to remove targeted parasitized cells.

[IEDB](#)'s analysis resource tool was used to identify MHC II restricted epitopes (HTL). The fasta sequence of the protein will be entered in this MHC-II binding predictions tool. The prediction method selected was NN-align 2.3 and the MHC resource specie was human and locus was HLA-DR. All the HLA-DRB1* alleles were selected and the predicted length of epitope was 15 [23] [27].

3.5.3 Antigenicity, Allergenicity, Toxicity and Sub-Cellular Localization Profiling of Epitopes

The predicted T-Cell epitopes undergo different selection criteria. Various prediction parameters were applied to estimate the eligibility of epitopes be a part of vaccine construct. Antigenicity, allergenicity, trans-membrane helices prediction were iniatially used as selection parameters. [VaxiJen v2.0](#) at threshold 0.4 was used to determined the peptides antigenicity. Epitopes with antigenicity 0.4 or above were selected.

Allergenicity of the antigenic epitopes was predicted through [AllerTOP v2.0](#) [27]. The non-allergen peptides were selected. The sub-cellular localization of non-allergen epitopes were determined using [TMHMM-2.0](#) was used [28]. This will also discriminate between soluble and membrane proteins. The peptides which were located outside were selected. [ToxinPred](#) was employed to determine the toxicity of the selected epitopes. Non-toxin epitopes were selected.

3.5.4 Determination of Physiochemical Properties and other Vital Features

Other physiochemical properties of non-toxic epitopes including hydrophobicity, hydrophilicity, charge, theoretical isoelectric point value (PI) were also determined using ToxinPred. [ProtParam](#) server was used to determine the molecular weight of the non-toxin epitopes.

3.6 Epitope Conservation Analysis

The epitope conservancy analysis of B-cells, MHC class-I and MHC class-II alleles was performed by using [IEDB](#) conservancy tool. This will predict the percentage of sequences matches at identity greater than or equal to 100 %.

3.7 Population Coverage Prediction of T-Cell Epitopes

Proportion of individuals predicted to respond to a given set of epitopes with known MHC restrictions were calculated by the population coverage analysis tool. The HLA genotypes frequencies among different populations of the world vary [24]. Difference in transmission and expression of the HLA alleles helps in designing epitope-based vaccine.

Thus, the binding ability of a specific epitope selected for the multivalent vaccine to HLA-I and II molecules is affected by HLA polymorphism [30]. Therefore, in this study it was critical to find out the extent that how much the designed multivalent vaccine constructs covers world population.

So, for determination of the population coverage by T cell (CTL and HTL) epitopes, the population coverage analysis tool of [IEDB](#) was used.

3.8 Construction of Multivalent Vaccine Design

On the basis of high antigenicity, high binding affinity (low IC₅₀ value) and non-allergenic nature, a set of B-Cell (LBL), and T-cell including CTL (MHC-I) and HTL (MHC-II) epitopes were selected. The multivalent vaccine was designed by joining the adjuvant, epitopes of T- (HTL and CTL) and B-cell with their respective linkers [27].

The selected epitopes were linked together through help of linkers for designing multi-vaccine construct. Linkers help to enhance the epitopes presentation and separated them properly. Linkers are important because of two reasons; (1) to avoid junctional epitope (neo-epitope) formation by separating epitopes, (2) and to improve epitope presentation. The linkers used in multivalent vaccine construct were EAAK, GPGPG and AAY [24].

The addition of an adjuvant at the N-terminus of vaccine construct with the help of EAAK linker. The adjuvant helps to produce a stronger immune response of vaccine in people receiving that vaccine to protect them from certain disease. The Cholera enterotoxin subunit B was added as an adjuvant at the N-terminus of vaccine with the use of EAAK linker.

CTB is linked with chemical or genetic manipulations and antigens and non-toxic component of cholera toxin provides a strongly enhanced immune response. CTB due to its versatile nature can influence immune responses in both directions and thus, makes it a promising adjuvant for vaccine development [25].

The EAAK linker separates the CTB from other vaccine domains, thus minimizes the interaction. After the adjuvant, B-Cell epitopes were linked. To connect B-Cell epitopes with each other, KK linker was used. B-cell epitopes were then linked to HTL (MHC-II epitopes) by GPGPG linker. HTLs were connected with each other by GPGPG linker. To link, HTL with CTL (MHC-I epitopes), AAY linker was used. Also CTLs were connected with each other by AAY linker. At last, the His-tag (six histamines) was added at the C-terminus of vaccine to finish the joining of the vaccine final construct [23].

3.8.1 Antigenicity and Allergenicity Profiling of Vaccine Construct

Analysing antigenicity is important step in designing vaccine. The vaccine antigenicity was predicted through [VaxiJen v2.0](#) tool at threshold 0.4. This tool generates antigenic score of vaccine protein sequence based on physiochemical properties of vaccine construct. If the vaccine construct has antigenic value above 0.4, then it will be selected for further analysis [28] [30].

Online prediction server was used to make sure that the vaccine does not cause any allergic reaction. To determine the allergenicity of the vaccine protein, and [AllertTOP v2.0](#) server was used, which predicts the allergenic potential of vaccine. If the protein is non-allergn, it will be selected [24].

3.8.2 Physicochemical Properties and Solubility Prediction of Vaccine Construct

[ProtParam](#) tool an online freely accessible web server, was applied to determine the physiochemical properties and molecular weight (MW) of vaccine, its theoretical isoelectric point value (PI), its instability index (II), its aliphatic index and GRAVY (grand average of hydropathicity) [24]. [SOLUPROT v1.0](#) is used to find the solubility of vaccine construct. This tool predicts vaccine expression in organism E.coli. If the vaccine solubility is above 0.5, the protein will have good expression in humans.

3.9 Structure Prediction, Refinement and Validation of Multi-epitope Vaccine

The secondary structure and tertiary structure was predicted. The validation and refinement of vaccine 3D structure was done. For this purpose, following steps were considered.

3.9.1 Prediction of Secondary Structure

[PSIPRED](#) server was used for the secondary structure prediction of vaccine construct. This server shows the results with high accuracy.

3.9.2 Prediction of Vaccine 3D Structure

The tertiary structure of vaccine was predicted using [iTASSER](#) server. This tool generated 3D model of protein from the sequence of its amino acids. To predict the accuracy of protein model, iTASSER also provides confidence score. Pymol software is used to view and check the 3D (tertiary) structure.

3.9.3 Refinement of Vaccine 3D Structure

The predicted tertiary structure model was refined by using [GalaxyRefine](#) and [trRosetta](#). These tools improved the global and structural quality of tertiary structure [24].

3.9.4 Validation of Vaccine 3D Structure

The additional validation of the protein vaccine structure was obtained through Ramachandran Plot by using [Ramachandran plot](#) server. The percentage of error comparisons of predicted structure residues of the refined protein compared to unrefined protein was predicted using [ERRAT](#) server [23]. ERRAT validated refined 3D structure of protein. The errat value before and after refinement was predicted [22]. The Z-score of refined structure was predicted by [molprobit](#) server.

3.9.5 Proteasomal Cleavages Prediction of MHC-ligands

[NetChop-3.1](#) at threshold 0.5 was used for the prediction of sites of proteasomal cleavages.

The pathway processing of MHC-I ligands mostly composed of two steps:

1. Initially proteasome degrades the protein,
2. The degraded products were transported along with antigen processing (TAP) to the endoplasmic reticulum via associated transporter.

In the endoplasmic reticulum, MHC-I molecules first bounded the antigen peptides and then presented them on cell surface. The fasta sequence of the vaccine protein was entered at threshold 0.5 to predict the number of cleavage sites [24].

3.10 Molecular Docking with TLR2

The immunologic receptors including toll-like receptors (TLRs) are necessary for eliciting response of immune system against infection caused by pathogenic organisms. During infection caused by *S.gordonii*, TLR involved in evoking immune response in humans is TLR-2 [5].

Protein-peptide molecular docking technique was used to determine the best binding mode of the multivalent vaccine construct to TLR-2 [29]. [RCSB](#) database retrieves the protein sequence of TLR-2. After retrieving the protein sequence, docking of the ligand and receptor was done. [ClusPro](#) server was used to carry out docking on TLR-2 (receptor) and vaccine construct (ligand). The interactions between the ligand (vaccine) and receptor was determined through [PDBePISA](#).

3.11 Immune-simulation

C-IMMSIM, an online web-server was used for immune simulation of vaccine construct. It simulates the response of host immune system to the multivalent vaccine design construct. The server is based on modeling approach and estimated the effect caused by foreign particle or antigen on the immune system using PSSM method. C-IMMSIM calculated the production of cytokines, interferon and antibodies after the injection of vaccine [22].

3.12 Gene Cloning

3.12.1 Sequence Translation of Vaccine Protein and Codon Adaptation

[EMBOSS Backtranseq webserver](#) was used to translate the protein of the vaccine. The server result shows that the protein was translated to nucleotides. The vaccine nucleotide sequence from EMBOSS Backtranseq was entered into [JCat](#) to adapt codon usage of vaccine [24].

In codon-adaptation, the optimized DNA sequence was obtained by selecting *Escherichia coli* (Strain K12) as an organism. Besides this, the DNA/RNA sequence option was selected because the protein sequence was translated. Additional options selected were (i) avoid prokaryotic ribosome binding sites (ii) avoid Restriction Enzymes cleavage sites (iii) avoid rho-independent transcription terminators.

These were used to avoid uncontrolled transcription and to avoid any change in nucleotide residues after codon adaptation for optimum cloning and expression. After the optimization of codon, the multivalent vaccine polypeptide with improved DNA sequence was obtained [22].

3.12.2 In-silico Cloning through SnapGene

The improved DNA sequence predicted from process of codon adaptation was subjected to in-silico cloning through [SnapGene](#) tool. The improved codon has been inserted into multiple cloning sites (MCS) of pET-28a (+) of the *E. coli* vector [22].

3.13 Overview of Methodology

Figure given below is the summary of methodology that was used to achieved objective of the research project.

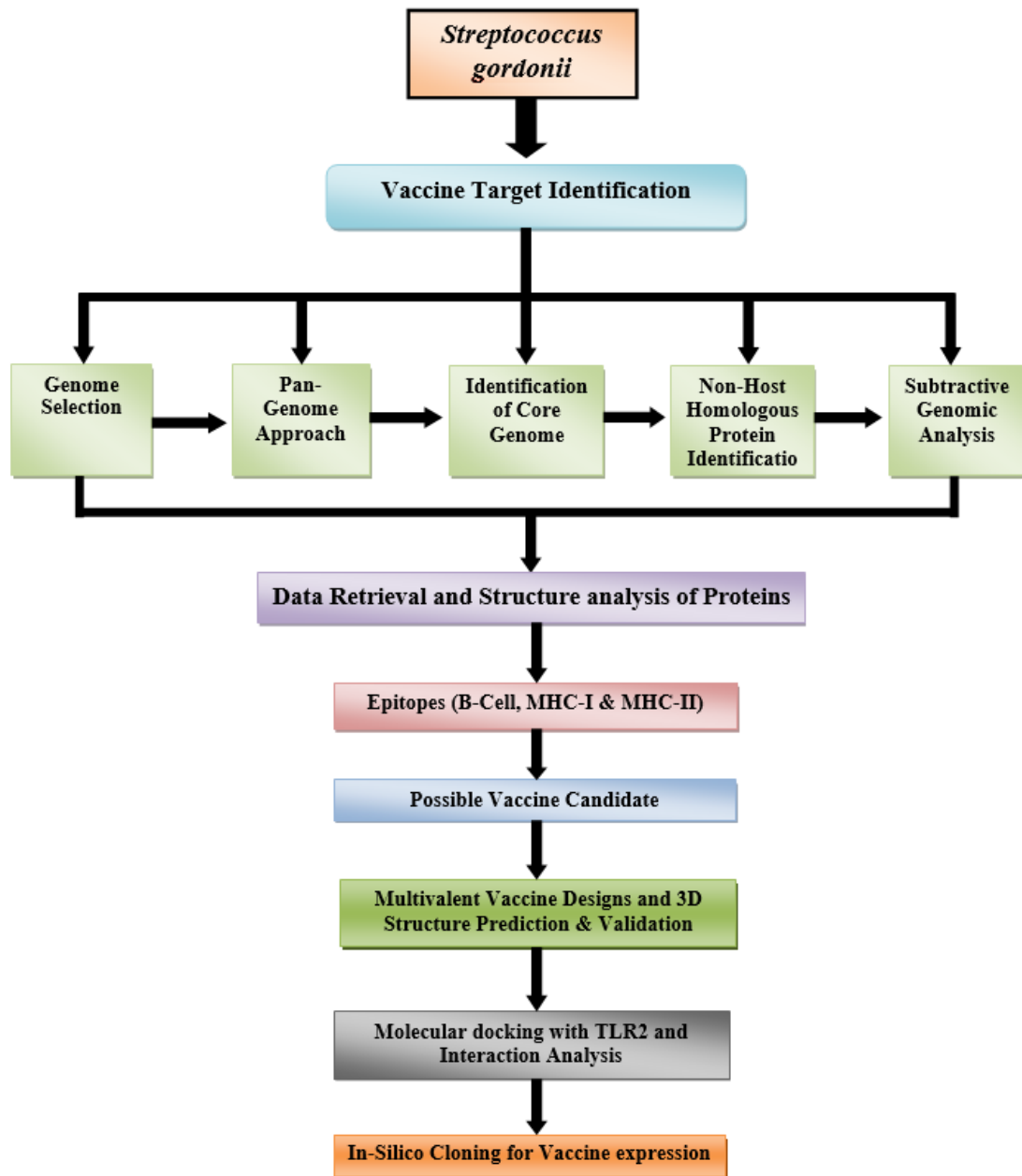


FIGURE 3.1: Flowchart of methodology

Chapter 4

Results and Discussion

4.1 Vaccine Target Prediction

4.1.1 Sequence Retrieval and Genome Selection

Genome Assembly and Annotation report of *S. gordonii* retrieved from NCBI shows total 91 strains of *Streptococcus gordonii*. Out of 91 strains of *S.gordonii* 13 are complete, 65 are contig, and 13 are scaffold. The thirteen strains (complete) of *S. gordonii* are included in this study that has complete genome. The proteins and genes of the 13 complete genomes of *S. gordonii* were retrieved from National Center for Biotechnology Information NCBI.

4.1.2 Identification of Core Genome by Pan-genome Approach

The core genome of *S. gordonii* was identified using EDGAR 2.3 software by pan genome approach. The EDGAR software requires selection of reference strain. The reference strain selected was “Challis substr. CH1” based on its release date that was compared with other 12 strains. The core genomes that were common in all the strains were selected. The total number of genes identified in pan-genome is 2835 genes; out of these 1225 were core genes.

4.1.3 Homologous and Non-homologous Protein Identification by using Subtractive Genomic Analysis

Sequence between the host and pathogen proteome are subtracted to identify proteins important for pathogen but not present in host by using subtractive genomics. These non-homologous genes were important for the survival of an organism and are called essential genes.

NCBI-BLASTp was used to identify non-homologous protein. FASTA sequence of all proteins was obtained from NCBI and then they were BLAST against humans to identify homologous and non-homologous proteins. From 1255 core genes, 643 were identified as non-homologous proteins.

4.1.4 Identification of Essential Genes

The identified non-host homologous proteins were used to identify essential proteins by inserted the proteins in DEG database: database of essential genes to find out the essential proteins by using thresholds. The default parameters will be selected, E value = 0.0001, bit score > 100, and identity > 25%. A total of 20 proteins were attained.

4.2 Collection of Surface Proteins

From the 20 essential proteins five surface proteins present in *S. gordonii* were selected for designing vaccine against *S. gordonii*. The subcellular localization of the proteins was identified using CELLO.

The five surface proteins include two extracellular proteins (YSIRK_signal domain protein, and Peptidase C51 domain-containing protein) and three membrane proteins (AraC family transcriptional regulator, Glycosyl transferase, Peptidoglycan D, D-transpeptidase FtsI).

TABLE 4.1: List of Proteins based on their Sub-Cellular Localization

Sr. #	Protein Name	Subcellular Localization
1	YSIRK_signal domain protein	Extracellular/cell wall
2	Peptidase C51 domain-containing protein	Extracellular
3	AraC family transcriptional regulator	Membrane
4	Glycosyl transferase	Membrane
5	Peptidoglycan D,D-transpeptidase FtsI	Membrane

4.2.1 Complete Sequence Retrieval of Proteins

The amino acids sequence of all the five surface proteins was retrieved from Uniprot using uniprot id of the protein i.e., A0A0A6S1K7, Q9AB82, F0HZA2, Q9AEU1, D0CCM7 to design multivalent vaccine for immunity against infection caused by *S.gordonii*.

The gene name was also predicted from uniprot. The virulence of the protein targets was identified using VFDB. Molecular weight of the five collected proteins was calculated from ProtParam.

TABLE 4.2: Uniprot ID, Virulence and Molecular Weight of Proteins

Sr. #	Gene Name	Protein Name	Uniprot ID	Virulence	MW
1	NC01_06050	YSIRK_signal domain protein	A0A0A-6S1K7	Yes	5399 2.06
2	CC_0349	Peptidase C51 domain-containing protein	Q9AB82	Yes	2830 5.13
3	HMPREF 9381_0242	AraC family transcriptional regulator	F0HZA2	Yes	3401 2.54
4	gtf3 (nss)	Glycosyl transferase	Q9AEU1	Yes	3800 7.40

		Peptidoglycan			6765
5	ftsI	D,D-transpeptidase FtsI	D0CCM7	Yes	9.05

4.2.2 Screening Antigenicity and Allergenicity *S. gordonii* Proteins and Selection of Protein

VaxiJen v2.0 is a web server was used to predict the antigenicity of all the five proteins at thresholds 0.4. By submitting the FASTA sequences of the protein to VaxiJen v2.0 serve, the antigenicity of proteins will be predicted. The proteins whose antigenicity was above 0.4 were selected for further analysis. The results from the server showed that two proteins possesses antigenic properties namely YSIRK - signal domain protein and Peptidoglycan D,D-transpeptidase FtsI and these two proteins were selected for further analysis. Antigenic score revealed by this server shows 0.5118 for YSIRK-signal domain protein and 0.6184 for Peptidoglycan D,D-transpeptidase FtsI, so these were antigenic proteins while other three proteins were non-antigenic as these shows antigenic score less than 0.4. AllerTOP server is used to predict the allergenicity of the surface proteins. The result from the server shows that all the five proteins were non-allergen. The proteins that were antigenic and non-allergen proteins i.e., YSIRK-signal domain protein and Peptidoglycan D,D-transpeptidase FtsI were selected for further analysis.

TABLE 4.3: Antigenicity and Allergenicity of *S. gordonii* Surface Proteins

Sr.#	Protein Name	Uniprot ID	Antigenicity	Allergenicity
1	YSIRK-signal domain protein	A0A0A-6S1K7	0.5118	Non-allergen
2	Peptidase C51 domain-containing protein	Q9AB82	0.3556	Non-allergen
3	AraC family transcriptional regulator	F0HZA2	0.3336	Non-allergen
4	Glycosyl transferase	Q9AEU1	0.3780	Non-allergen

5	Peptidoglycan D,D-transpeptidase FtsI	D0CCM7	0.6184	Non-allergen
---	--	--------	--------	--------------

4.2.3 Structure Analysis by Screening of Physio-chemical Properties of Selected Proteins

Two selected proteins including YSIRK_signal domain protein and Peptidoglycan D, D-transpeptidase FtsI were subjected to structural analysis. The structure analysis of both the selected antigenic proteins will be done by screening physio-chemical properties of the antigenic proteins by using ProtParam tool (ExPASy).

YSIRK_signal domain protein has 469 amino acids with a molecular weight of 53992.06kDa while Peptidoglycan D, D-transpeptidase FtsI has 610 amino acids with a molecular weight of 67659.05kDa. The theoretical pI value shows the nature of the protein.

If the theoretical pI value of the protein is below 7, the nature of protein will be negative and if the theoretical pI value of the protein is above 7, the nature of protein will be positive.

The theoretical pI value of YSIRK_signal domain protein 9.36 and the theoretical pI value of Peptidoglycan D, D-transpeptidase FtsI 9.78. The theoretical pI values of both proteins were above 7, so the nature of both proteins was positive.

Other physiochemical properties including instability index, aliphatic index, gravy, half-life, atomic composition, the total number of positively charged residues (Arg + Lys) and the total number of negatively charged residues (Asp + Glu), were also predicted using protparam and shown in the table.

The SoluProt was used to predict solubility of the proteins. The result from SoluProt shows YSIRK_signal domain protein 0.547 and Peptidoglycan D, D-transpeptidase FtsI 0.537. DIANNA 1.1 was used to predict the disulphide bond in both proteins. The result shows that there was no disulfide bond in YSIRK_signal domain protein neither in Peptidoglycan D, D-transpeptidase FtsI.

TABLE 4.4: Physio-chemical Properties and Solubility of YSIRK_signal domain protein and Peptidoglycan D, D-transpeptidase FtsI

Property	YSIRK_signal domain protein	Peptidoglycan D,D-transpeptidase FtsI
Molecular weight	53992.06kDa	67659.05kDa
Number of amino acids	469	610
Theoretical pI value	9.36	9.78
Instability index	44.23	37.64
Aliphatic index	44.73	89.69
Gravy	-1.277	-0.355
	30 hours (mammalian reticulocytes, in vitro)	30 hours (mammalian reticulocytes, in vitro)
Estimated half life	>20 hours (yeast, in vivo)	>20 hours (yeast, in vivo)
	>10 hours (Escherichia coli, in vivo)	>10 hours (Escherichia coli, in vivo)
Atomic composition (Carbon, Hydrogen, Nitrogen, Oxygen, Sulfur)	C 2401 H 3676 N 686 O 730 S 5	C 3002 H 4855 N 857 O 877 S 21
Total number of negatively charged residues (Asp + Glu)	66	55
Total number of positively charged residues (Arg + Lys)	79	81
Protein solubility	0.547	0.537

4.2.4 Function and Pathway Analysis of Selected Protein

The function and pathway of selected proteins were retrieved from uniprot by using uniprot ID of the protein. The function of the protein includes its molecular and biological function. The function of Peptidoglycan D, D-transpeptidase FtsI is to catalyze cross-linking of the peptidoglycan cell wall at the division septum. The pathway of the protein was also predicted using KEGG web tool.

TABLE 4.5: Function and Pathway Analysis of YSIRK_signal Domain Protein and Peptidoglycan D, D-transpeptidase FtsI

Protein Name	Uniprot ID	Molecular Function	Biological Function	Pathway
YSIRK_signal domain protein	A0A0A6S1K7	Carbohydrate binding, Glycopeptides alpha-N-acetylgalactosaminidase activity		No hits
Peptidoglycan D, D-transpeptidase FtsI	D0C7CM7	Penicillin binding, peptidoglycan glycosyltransferase activity, serine-type D-Ala-D-Ala carboxypeptidase activity.	Cell wall organization, division septum assembly, FtsZ-dependent cytokinesis, peptidoglycan biosynthetic process, proteolysis, regulation of cell shape.	Cell wall biogenesis; peptidoglycan biosynthesis.

4.3 B-cell Epitope Prediction and Selection

B-cell epitopes obtained through ABCpred tool [24]. By entering the fasta sequence of each protein in the ABCpred tool, the B-cell epitopes of the antigenic

proteins were obtained. Selected length of epitope was 16 at threshold 0.5. Subsequently, 45 epitopes were predicted for YSIRK_signal domain-containing protein and 65 epitopes were predicted for Peptidoglycan D, D-transpeptidase FtsI.

4.3.1 Antigenicity, Allergenicity, Molecular Weight and Subcellular Localization Profiling of Epitopes

The physiochemical properties of the epitopes were also predicted. These physiochemical properties include antigenicity, allergicity, toxicity and subcellular localization of epitopes. The antigenicity was predicted by using VaxiJen v2.0 tool at threshold 0.4. AllerTOP v2.0 server was used to predict the allergenic potential of antigenic epitopes.

Non-allergen epitopes were selected. ProtParam tool an online freely accessible web server, was applied to determine the molecular weight (MW) of epitopes. ToxinPred determined the toxicity of the epitopes. Non-toxin epitopes was selected. The necessary peptides were selected based on high antigenic value. The peptides that were non-allergen, non-toxin, and present outside and having high antigenic value were selected, one epitope from YSIRK_signal domain protein was selected and two epitopes from Peptidoglycan D, D-transpeptidase FtsI were selected.

TABLE 4.6: B cell epitopes of Selected Protein via ABCpred

Protein	Sequence/ Epitope	Position	Score
YSIRK_ signal domain- containing protein	LWTPNGLT KGNENNAP	120	0.83
Peptidoglycan D,D-transpe- ptidase FtsI	FDMWRFYL LWAVVLLC EPGENITL SIDSRLQY	24 255	0.71 0.62

TABLE 4.7: Antigenicity, Allergicity, Toxicity, and Sub-cellular Localization Evaluation of B-Cell Epitopes

Protein	Sequence/ Epitope	Antige- nicity	Allerg- enicity	Toxi- city	Sub-cell- ular loca- lization	MW
YSIRK_signal domain-contai- ning protein	LWTPN	0.4168	Non- allergen	Non- toxic	Outside	172
	GLTKG					5.88
	NENNAP					
Peptidoglycan D,D-transpep- tidase FtsI	FDMWR	1.5824	Non- allergen	Non- toxic	Outside	207
	FYLLWA					5.56
	VVLLC					
	EPGENIT	1.0752	Non- allergen	Non- toxic	Outside	1835
	LSIDSR					
	LQY					

4.4 T-cell Epitope Selection

The epitopes of T-cell were predicted using T-cell epitope prediction tool of IEDB. T-cell epitopes includes MHC-I and MHC-II binding epitopes. The epitopes that were restricted to MHC-I molecules and MHC-II molecules were selected based on allergicity, antigenicity, IC50 value and other physiochemical properties [23].

4.4.1 Screening and Selection of MHC-I Restricted Epitopes

MHC-I restricted epitopes obtained using IEDB. Fasta sequence of the protein will be entered in this MHC-I binding predictions tool. The prediction method selected was ANN 4.0, the MHC resource specie was human and the predicted length of epitope was 9. All HLA-A* alleles were obtained. Approximately, 5520 alleles for 335 common peptides of YSIRK_signal domain-containing protein and

7200 alleles for 450 common peptides of Peptidoglycan D, D-transpeptidase FtsI were obtained. The significant epitopes from common epitopes were selected on the basis low IC50 value of the predicted peptides. The epitopes were selected on the basis of certain properties including determination of antigenicity, allergenicity, trans-membrane helicase and toxicity. The peptides with antigenicity above 0.4 were selected using VaxiJen v2.0.

TABLE 4.8: MHC-I alleles binding peptides of YSIRK_{signal domain}-containing protein and Peptidoglycan D,D-transpeptidase FtsI computed via IEDB along with their antigenicity computed via VaxiJen 2.0

Protein	Peptide Sequence/ Epitope	MHC-I alleles	ic50	Antig- enicity
YSIRK _{signal domain} protein	FYYPP	HLA-A*29:02, HLA-A*23:01,	63.24	1.9204
	FPDM	HLA-A*02:06, HLA-A*24:02, HLA-A*30:02, HLA-A*30:01, HLA-A*68:02, HLA-A*31:01, HLA-A*02:01, HLA-A*26:01, HLA-A*25:01, HLA-A*03:01, HLA-A*01:01, HLA-A*11:01, HLA-A*68:01, HLA-A*32:01		
	REPFY	HLA-A*32:01, HLA-A*24:02,	257.72	1.3015
	YPPF	HLA-A*23:01, HLA-A*02:06, HLA-A*30:01, HLA-A*29:02, HLA-A*30:02, HLA-A*26:01, HLA-A*31:01, HLA-A*03:01, HLA-A*01:01, HLA-A*25:01, HLA-A*11:01, HLA-A*68:02, HLA-A*02:01, HLA-A*68:01		
	VQVDS	HLA-A*02:06, HLA-A*26:01,	739.1	1.4495
	VTEE	HLA-A*11:01, HLA-A*31:01, HLA-A*30:01, HLA-A*02:01, HLA-A*30:02, HLA-A*29:02,		

		HLA-A*01:01, HLA-A*68:01, HLA-A*23:01, HLA-A*25:01, HLA-A*68:02, HLA-A*03:01, HLA-A*32:01, HLA-A*24:02		
Peptid- oglycan D, D-trans- peptidase FtsI	WAVV LLCFV	HLA-A*02:06, HLA-A*68:02, HLA-A*02:01, HLA-A*30:01, HLA-A*31:01, HLA-A*01:01, HLA-A*68:01, HLA-A*30:02, HLA-A*29:02, HLA-A*26:01, HLA-A*23:01, HLA-A*03:01, HLA-A*25:01, HLA-A*11:01, HLA-A*24:02, HLA-A*32:01	14.9	1.2845
	VLLC	HLA-A*02:01, HLA-A*02:06,	21.9	1.5192
	FVVLI	HLA-A*32:01, HLA-A*23:01, HLA-A*30:01, HLA-A*68:02, HLA-A*31:01, HLA-A*24:02, HLA-A*29:02, HLA-A*03:01, HLA-A*01:01, HLA-A*11:01, HLA-A*68:01, HLA-A*30:02, HLA-A*26:01, HLA-A*25:01		
	LWAV	HLA-A*23:01, HLA-A*24:02,	45.65	1.5198
	VLLCF	HLA-A*29:02, HLA-A*02:06, HLA-A*30:02, HLA-A*02:01, HLA-A*31:01, HLA-A*30:01, HLA-A*01:01, HLA-A*32:01, HLA-A*03:01, HLA-A*26:01, HLA-A*68:02, HLA-A*25:01, HLA-A*68:01, HLA-A*11:01		
	AVVL	HLA-A*02:06, HLA-A*02:01,	80.01	1.4047
	LCFVV	HLA-A*68:02, HLA-A*30:01, HLA-A*31:01, HLA-A*11:01, HLA-A*32:01, HLA-A*30:02,		

HLA-A*03:01, HLA-A*29:02,
 HLA-A*26:01, HLA-A*01:01,
 HLA-A*23:01, HLA-A*24:02,
 HLA-A*68:01, HLA-A*25:01

4.4.1.1 Antigenicity, Allergenicity, Molecular Weight and Sub-cellular Localization Profiling

The antigenicity, allergicity, toxicity and subcellular localization of epitopes were also predicted. The necessary peptides were selected based on high antigenic value. Antigenicity analysis is important step in designing vaccine. The antigenicity of epitopes was predicted by using VaxiJen v2.0 tool at threshold 0.4. The epitopes having antigenic value above 0.4 were selected for further analysis. Online prediction server was used to make sure that the vaccine does not cause any allergic reaction. For this, AllerTOP v2.0 tool to predict allergenicity.

TABLE 4.9: Antigenicity, Allergenicity, Toxicity, and Sub-cellular Localization of Selected MHC-I Restricted Epitopes

Protein	Sequence/ Epitope	Antig- enicity	Allerg- enicity	Toxi- city	Sub-cellular localization
YSIRK_signal domain-contai- ning protein	FYYPPF	1.9204	Non-	Non-	Outside
	PDM		allergen	toxic	
	REPFY	1.3015	Non-	Non-	Outside
	YPPF		allergen	toxic	
	VQVDS	1.4495	Non-	Non-	Outside
VTEE	allergen		toxic		
Peptidoglycan	WAVVL	1.2845	Non-	Non-	Outside
	LCFV		allergen	toxic	
D,D-transpep- tidase FtsI	VLLCF	1.5192	Non-	Non-	Outside
	VVLI		allergen	toxic	
	LWAVV	1.5198	Non-	Non-	Outside
LLCF	allergen		toxic		

AVVLL	1.4047	Non-	Non-	Outside
CFVV		allergen	toxic	

The epitopes that were non-allergen were selected and other allergen epitopes were discarded. Subcellular localization of the non-allergen epitope was then predicted using TMHMM-2.0 tool. The epitopes that were present outside were selected and other will be discarded.

ToxinPred was then used to predict the toxicity of the epitopes that were present outside. Non-toxin peptides will be selected. After applying the above tool of prediction, only 37 peptides out of 335 peptides of YSIRK_signal domain-containing protein and 54 peptides out of 450 peptides of Peptidoglycan D,D-transpeptidase FtsI were left that were antigenic, non-allergen, non-toxin and present outside.

The epitopes were arranged according to their low to high IC₅₀ value. Three epitopes of YSIRK_signal domain-containing protein and four epitopes of Peptidoglycan D,D-transpeptidase FtsI were selected on the basis of their properties.

4.4.1.2 Physiochemical Properties and Vital Features Determination

Physiochemical properties including hydrophobicity, hydrophilicity, charge and pI value of the selected epitopes were predicted through ToxinPred. ProtParam tool an online freely accessible web server, was applied to determine the molecular weight (MW) of vaccine. Another vital feature is the prediction of peptide digesting enzymes by using Peptide cutter tool.

TABLE 4.10: Physio-Chemical Properties of MHC-I Restricted Epitopes

Protein	Sequence/ Epitope	Hydrop- hobicity	Hydrop- hilicity	Charge	pI	MW
YSIRK_signal domain-con taining protein	FYYPP	0.07	-0.88	-1	3.8	117
	FPDM					6.35
	REPF	-0.15	-0.4	0	6.35	121

	YYPPF					5.37
	VQVD	-0.16	0.51	-3	3.58	100
	SVTEE					5.05
Peptidoglycan	WAVV	0.44	-1.72	0	5.85	104
D, D-transpep-	LLCFV					9.34
tidase FtsI	VLLCF	0.51	-1.69	0	5.85	101
	VVLI					8.37
	LWAV	0.44	-1.76	0	5.85	106
	VLLCF					3.37
	AVVLL	0.46	-1.51	0	5.85	96
	CFVV					2.26

If Peptides are digested by fewer enzymes, they are considered to be stable peptides and are more favorable vaccine targets whereas the peptides that are digested by several enzymes are considered to be non-stable. The result from the peptide cutter tool shows that the peptides are stable because these are digested by enzymes.

TABLE 4.11: Non-Digesting Enzymes of MHC-I Restricted Epitopes

Protein	Sequence/ Epitope	Non-digesting Enzymes
YSIRK_sig- gnal domain- containing protein	FYYPPFPDM	Arg-C proteinase, BNPS-Skatole, Caspase 1-10, Clostripain, Enterokinase, Factor Xa, Glutamyl endopeptidase, GranzymeB, Hydroxylamine, Iodosobenzoic acid, LysC, LsyN, NTCB (2-nitro-5-thiocyanobenzoic acid), Proline-endopeptidase, Staphylococcal peptidase I, Thermolysin, Thrombin, Tobacco etch virus protease, Trypsin
	REPFYYPPF	Asp-N endopeptidase, BNPS-Skatole, CNBr, Caspase1-10, Enterokinase, Factor Xa, Formic acid, GranzymeB, Hydroxylamine, Iodosobenzoic acid, LysC,

		LsyN, NTCB (2-nitro-5-thiocyanobenzoic acid), Pepsin (pH1.3), Pepsin (pH>2), Proline-endopeptidase, Thrombin, Tobacco etch virus protease
	VQVDSVTEE	Arg-C proteinase, BNPS-Skatole, CNBr, Caspase1-10, Chymotrypsin-high specificity (C-term to [FYW], not before P), Chymotrypsin-low specificity (C-term to [FYWL], not before P), Clostripain, Enterokinase, Factor Xa, GranzymeB, Hydroxylamine, Iodosobenzoic acid, LysC, LsyN, NTCB (2-nitro-5-thiocyanobenzoic acid), Pepsin (pH1.3), Pepsin (pH>2), Proline-endopeptidase, Thrombin, Tobacco etch virus protease, Trypsin
Peptido- glycan D,D- transpe - ptidase FtsI	WAVVLLCFV	Arg-C proteinase, Asp-N endopeptidase, Asp-N endopeptidase + N-terminal Glu, CNBr, Caspase1-10, Clostripain, Trypsin, Enterokinase, Factor Xa, Formic acid, Glutamyl endopeptidase, GranzymeB, Hydroxylamine, LysC, LsyN, Proline-endopeptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease,
	VLLCFVCLI	Arg-C proteinase, Asp-N endopeptidase, Asp-N endopeptidase + N-terminal Glu, BNPS-Skatole, CNBr, Caspase1-10, Clostripain, Enterokinase, Factor Xa, Formic acid, Glutamyl endopeptidase, GranzymeB, Hydroxylamine, Thrombin, Iodosobenzoic acid, LysC, LsyN, Proline-endopeptidase, Staphylococcal peptidase I, Tobacco etch virus protease, Trypsin

LWAVVLLCF	Arg-C proteinase, Asp-N endopeptidase, Asp-N endopeptidase + N-terminal Glu, CNBr, Caspase1-10, Clostripain, Enterokinase, Factor Xa, Formic acid, Glutamyl endopeptidase, GranzymeB, Hydroxylamine, Iodosobenzoic acid, LysC, LsyN, Proline-endopeptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease, Trypsin
AVVLLCFVV	Arg-C proteinase, Asp-N endopeptidase, Asp-N endopeptidase + N-terminal Glu, BNPS-Skatole, CNBr, Caspase1-10, Clostripain, Enterokinase, Factor Xa, Formic acid, Glutamyl endopeptidase, GranzymeB, Hydroxylamine, Thrombin, Iodosobenzoic acid, LysC, LsyN, Proline-endopeptidase, Staphylococcal peptidase I, Tobacco etch virus protease, Trypsin

4.4.2 Screening of MHC-II Restricted Epitopes

MHC II restricted epitopes were obtained using IEDB. The fasta sequence of the protein will be entered in this MHC-II binding predictions tool. The prediction method selected was NN-align 2.3 and the MHC resource specie was human and locus was HLA-DR. All the HLA-DRB1* alleles were obtained and the predicted length of epitope was 15. Approximately, 4921 alleles for 257 common peptides of YSIRK_signal domain-containing protein and 6270 alleles for 300 common peptides of Peptidoglycan D,D-transpeptidase FtsI were obtained. The significant epitopes from common epitopes were selected on the basis low IC50 value of the predicted peptides. The epitope were selected on the basis of certain properties including determination of antigenicity, allergicity, trans-membrane helicase and

toxicity. VaxiJen v2.0 is used to predict antigenicity of epitopes. The peptides with antigenicity above 0.4 were selected.

TABLE 4.12: MHC-II Alleles Binding Peptides of YSIRK_signal domain-containing Protein and Peptidoglycan D,D-transpeptidase FtsI computed via IEDB along with their antigenicity computed via VaxiJen 2.0

Protein	Sequence/ Epitope	MHC-II alleles	ic50	Antig- enicity
YSIRK	APFVF	HLA-DRB1*04:01, HLA-DRB1*08:02,	30.9	1.12
_signal	KPEST	HLA-DRB1*16:02, HLA-DRB1*08:01,		14
domain-	PAPKL	HLA-DRB1*04:05, HLA-DRB1*01:01,		
protein		HLA-DRB1*10:01, HLA-DRB1*04:04, HLA-DRB1*11:01, HLA-DRB1*07:01, HLA-DRB1*09:01, HLA-DRB1*13:02, HLA-DRB1*15:01, HLA-DRB1*04:02, HLA-DRB1*13:01, HLA-DRB1*04:03, HLA-DRB1*01:03, HLA-DRB1*03:01, HLA-DRB1*12:01		
	PFVFKP	HLA-DRB1*04:01, HLA-DRB1*08:02,	33.1	1.35
	EKPEST	HLA-DRB1*16:02, HLA-DRB1*08:01,		79
	PAPKLD	HLA-DRB1*04:05, HLA-DRB1*01:01, HLA-DRB1*10:01, HLA-DRB1*04:04, HLA-DRB1*11:01, HLA-DRB1*07:01, HLA-DRB1*09:01, HLA-DRB1*13:02, HLA-DRB1*15:01, HLA-DRB1*04:02, HLA-DRB1*13:01, HLA-DRB1*01:03, HLA-DRB1*03:01, HLA-DRB1*12:01, HLA-DRB1*04:03		
	FVFKP	HLA-DRB1*04:01, HLA-DRB1*08:02,	39.8	1.40
	ESTPAP	HLA-DRB1*08:01, HLA-DRB1*16:02,		43
	KLDM	HLA-DRB1*01:01, HLA-DRB1*04:05, HLA-DRB1*10:01, HLA-DRB1*04:04, HLA-DRB1*07:01, HLA-DRB1*11:01,		

		HLA-DRB1*09:01, HLA-DRB1*13:02, HLA-DRB1*15:01, HLA-DRB1*04:02, HLA-DRB1*13:01, HLA-DRB1*01:03, HLA-DRB1*12:01, HLA-DRB1*03:01, HLA-DRB1*04:03		
Peptido- glycan- D,D-tra nspept- idase FstI	PGEN ITLSID SRLQY	HLA-DRB1*03:01, HLA-DRB1*13:02, HLA-DRB1*09:01, HLA-DRB1*04:04, HLA-DRB1*15:01, HLA-DRB1*04:05, HLA-DRB1*12:01, HLA-DRB1*13:01, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*01:03, HLA-DRB1*01:01, HLA-DRB1*04:01, HLA-DRB1*04:03, HLA-DRB1*11:01, HLA-DRB1*16:02, HLA-DRB1*10:01, HLA-DRB1*08:01, HLA-DRB1*04:02	8.2	1.07 22
	GTMA YGYGL NATILQ	HLA-DRB1*09:01, HLA-DRB1*04:01, HLA-DRB1*01:01, HLA-DRB1*04:02, HLA-DRB1*15:01, HLA-DRB1*10:01, HLA-DRB1*04:03, HLA-DRB1*04:05, HLA-DRB1*16:02, HLA-DRB1*13:02, HLA-DRB1*07:01, HLA-DRB1*12:01, HLA-DRB1*08:02, HLA-DRB1*11:01, HLA-DRB1*04:04, HLA-DRB1*08:01, HLA-DRB1*01:03, HLA-DRB1*13:01, HLA-DRB1*03:01	11.8	1.04 47
	VLLCFV VVLIAF AFYV	HLA-DRB1*04:03, HLA-DRB1*12:01, HLA-DRB1*01:03, HLA-DRB1*04:02, HLA-DRB1*07:01, HLA-DRB1*15:01, HLA-DRB1*01:01, HLA-DRB1*08:01, HLA-DRB1*13:01, HLA-DRB1*13:02, HLA-DRB1*11:01, HLA-DRB1*16:02, HLA-DRB1*09:01, HLA-DRB1*08:02,	26.3	0.91 42

	HLA-DRB1*10:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:04, HLA-DRB1*04:05		
AQIIG	HLA-DRB1*04:04, HLA-DRB1*04:05,	30.4	0.99
LTNSEG	HLA-DRB1*04:01, HLA-DRB1*04:03,		63
QGIE	HLA-DRB1*08:02, HLA-DRB1*10:01, HLA-DRB1*13:02, HLA-DRB1*07:01, HLA-DRB1*16:02, HLA-DRB1*01:01, HLA-DRB1*04:02, HLA-DRB1*12:01, HLA-DRB1*09:01, HLA-DRB1*15:01, HLA-DRB1*11:01, HLA-DRB1*13:01, HLA-DRB1*08:01, HLA-DRB1*01:03, HLA-DRB1*03:01		

4.4.2.1 Antigenicity, Allergenicity, Molecular Weight and Sub-cellular Localization Profiling

The antigenicity, allergicity, toxicity and subcellular localization of epitopes were also predicted. The necessary peptides were selected based on high antigenic value. Antigenicity analysis is important step in designing vaccine. The vaccine antigenicity was predicted by using VaxiJen v2.0 tool at threshold 0.4. This tool generates the antigenic score of the protein sequence of vaccine based on physiochemical properties of vaccine construct. If the vaccine construct has antigenic value above 0.4, then it will be selected for further analysis. VaxiJen v2.0 at threshold 0.4 is used to predict antigenicity of epitopes. After this, epitopes were subjected to AllerTOP v2.0 tool to predict allergenicity. The epitopes that were non-allergen were selected and other allergen epitopes were discarded. Transmembrane helix case of the non-allergen epitope were then predicted using TMHMM-2.0 tool. The epitopes that were present outside were selected and other will be discarded.

ToxinPred was then used to predict the toxicity of the epitopes that were present outside. Non-toxin peptides will be selected. After applying the above properties

only 31 peptides out of 257 peptides of YSIRK_signal domain-containing protein and 39 peptides out of 300 peptides of Peptidoglycan D,D-transpeptidase FtsI were left that were antigenic, non-allergen, non-toxin and present outside.

The epitopes that were antigenic, non-allergen, non-toxin, present outside and having low IC50 value were selected. Three epitopes of YSIRK_signal domain-containing protein and four epitopes of Peptidoglycan D,D-transpeptidase FtsI were selected on the basis of above properties.

TABLE 4.13: Antigenicity, Allergenicity, Toxicity, and Sub-cellular localization of selected MHC-I Restricted Epitopes

Protein	Sequence/ Epitope	Antig- enicity	Allerg- enicity	Toxi- city	Sub-cellular localization
YSIRK_signal domain-conta- ining protein	APFVFKP	1.1214	Non- allergen	Non- toxic	Outside
	ESTPAPKL				
	PFVFKPE				
Peptidoglycan D,D-transpep- tidase FtsI	STPAPKLD	1.3579	Non- allergen	Non- toxic	Outside
	FVFKPEST				
	PAPKLDM				
	PGENITLSI				
	DSRLQY				
Peptidoglycan D,D-transpep- tidase FtsI	GTMAYGY	1.0447	Non- allergen	Non- toxic	Outside
	GLNATILQ				
	VLLCFVVL				
	IARAFYV				
Peptidoglycan D,D-transpep- tidase FtsI	AQIIGLTN	0.9142	Non- allergen	Non- toxic	Outside
	SEGQGIE				
Peptidoglycan D,D-transpep- tidase FtsI	AQIIGLTN	0.9963	Non- allergen	Non- toxic	Outside
	SEGQGIE				

4.4.2.2 Determination of Physicochemical properties and Vital Features

Physicochemical properties including hydrophobicity, hydrophilicity, charge and pI value were predicted through ToxinPred. Molecular weights of the selected epitopes were also predicted through online freely accessible web server ProtParam.

Another vital feature is the prediction of peptide digesting enzymes by using Peptide cutter tool. If Peptides are digested by fewer enzymes, they are considered to be stable peptides and are more favorable vaccine targets whereas the peptides that are digested by several enzymes are considered to be non-stable.

The result from the peptide cutter tool shows that the peptides are stable because these are digested by small number of enzymes.

TABLE 4.14: Physio-chemical properties of MHC-II Restricted Epitopes

Protein	Sequence/ Epitope	Hydrop- hobicity	Hydrop- hilicity	Charge	pI	MW
YSIRK_ signal	APFVFKPE	-0.05	-0.03	1	8.94	162
	STPAPKL					8.93
domain- containing	PFVFKPES	-0.11	0.21	0	6.42	167
	TPAPKLD					2.94
protein	FVFKPEST	-0.09	0.12	0	6.42	170
	PAPKLDM					7.02
Peptido glycan D,D-	PGENITLS	-0.17	0.01	-1	4.38	170
	IDSRLQY					5.89
transpept- idase FtsI	GTMAYGY	0.09	-0.85	0	5.87	157
	GLNATILQ					2.8
	VLLCFVVL	0.3	-1.3	1	8.57	172
	IARAFYV					6.2
	AQIIGLTN	-0.02	-0.08	-2	3.8	152
	SEGQGIE					9.67

TABLE 4.15: Non-digesting enzymes of MHC-II Restricted Epitopes

Protein	Sequence/ Epitope	Non-digesting Enzymes
YSIRK_ signal	APFVFKP	Arg-C proteinase, Asp-N endopeptidase,
	ESTPAPKL	BNPS-Skatole, CNBr, Caspase1-10,
domain-		Clostripain, Enterokinase, Factor Xa,

containing protein		Formic acid, GranzymeB, Hydroxylamine, Iodosobenzoic acid, NTCB (2-nitro-5- thiocyanobenzoic acid), Thrombin, Tobacco etch virus protease
	PFVFKPE	Arg-C proteinase, BNPS-Skatole, CNBr,
	STPAPKLD	Caspase1-10, Clostripain, Enterokinase, Factor Xa, GranzymeB, Hydroxylamine, Iodosobenzoic acid, NTCB (2-nitro-5- thiocyanobenzoic acid), Thrombin, Tobacco etch virus protease
	FVFKPEST	Arg-C proteinase, BNPS-Skatole,
	PAPKLDM	Caspase1-10, Clostripain, Enterokinase, Factor Xa, GranzymeB, Hydroxylamine, Iodosobenzoic acid, NTCB (2-nitro-5- thiocyanobenzoic acid), Thrombin. Tobacco etch virus protease
Peptido- glycan D, D-trans- Peptid- ase FtsI	PGENITLS	BNPS-Skatole, CNBr, Caspase1-10,
	IDSRLQY	Enterokinase, Factor Xa, GranzymeB, Hydroxylamine, Iodosobenzoic acid, LysC, LsyN, NTCB (2-nitro-5- thiocyanobenzoic acid), Proline- endopeptidase, Thrombin. Tobacco etch virus protease
	GTMAYGY	Arg-C proteinase, Asp-N endopeptidase,
	GLNATILQ	Asp-N endopeptidase + N-terminal Glu, BNPS-Skatole, Caspase1-10, Clostripain, Enterokinase, Factor Xa, Formic acid, Glutamyl endopeptidase, GranzymeB, Hydroxylamine, Iodosobenzoic acid, LysC, LsyN, NTCB (2-nitro-5- thiocyanobenzoic acid), Proline- endopeptidase, Staphylococcal

	peptidase I, Thrombin. Tobacco etch virus protease, Trypsin
VLLCFVVL	Asp-N endopeptidase, Asp-N
IARAFYV	endopeptidase + N-terminal Glu, BNPS-Skatole, CNBr, Caspase1-10, Clostripain, Enterokinase, Factor Xa, Formic acid, Glutamyl endopeptidase, GranzymeB, Hydroxylamine, Iodosobenzoic acid, LysC, LsyN, Proline-endopeptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease
AQIIGLTN	Arg-C proteinase, Asp-N endopeptidase,
SEGQGIE	BNPS-Skatole, CNBr, Caspase1-10, Chymotrypsin-high specificity (C-term to [FYW], not before P), Clostripain, Enterokinase, Factor Xa, Formic acid, Glutamyl endopeptidase, GranzymeB, Hydroxylamine, Iodosobenzoic acid, LysC, LsyN, NTCB (2-nitro-5-thiocyanobenzoic acid), Proline-endopeptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease, Trypsin

4.5 Epitope Conservation Analysis

The epitope conservancy analysis of B-cells, MHC class-I and MHC class-II alleles was performed by using IEDB conservancy tool. This tool computes the degree of conservancy of an epitope within a given protein sequence set at a given identity level. Conservancy is defined as the fraction of protein sequences that contain the epitope, and Identity is the degree of correspondence (similarity) between two sequences.

TABLE 4.16: Conservation analysis of epitopes of YSIRK₂ signal domain protein computed via IEDB epitope conservancy analysis tool

Epitope Sequence	Epitope Length	Percentage of		
		Sequence matches at identity <= 100%	Minimum Identity	Maximum Identity
LWTPNG	16	0.00% (0/1)	31.25%	31.25%
LTKGNE				
NNAP				
FYYPP	9	0.00% (0/1)	33.33%	33.33%
FPDM				
REPFY	9	0.00% (0/1)	44.44%	44.44%
YPPF				
VQVD	9	0.00% (0/1)	55.56%	55.56%
SVTEE				
APFVFKPE	15	0.00% (0/1)	40.00%	40.00%
STPAPKL				
PFVFKPES	15	0.00% (0/1)	33.33%	33.33%
TPAPKLD				
FVFKPEST	15	0.00% (0/1)	33.33%	33.33%
PAPKLDM				

4.6 Population Coverage of Selected T-Cell Epitopes

Variation in the pattern of Human leukocyte antigen (HLA) is observed among geographical areas. IEDB tool was used to determine the T cell population coverage in vaccine construct. Result shows that the multi-epitope vaccine development strategy could cover 99.73% of the human population around the world and 96.84% in Pakistan.

The high coverage around the world and in Pakistani population indicates that our vaccine construct may work better around the world and also in Pakistani population.

TABLE 4.17: Conservation analysis of epitopes of Peptidoglycan D,D-transpeptidase FtsI computed via IEDB epitope conservancy analysis tool

Epitope Sequence	Epitope Length	Percentage of Sequence matches at identity \leq 100%	Minimum Identity	Maximum Identity
FDMWRFYL LWAVVLLC	16	0.00% (0/1)	31.25%	31.25%
EPGENITLS IDSRLQY	16	0.00% (0/1)	37.50%	37.50%
WAVVLLCFV VLLCFVCLI	9	0.00% (0/1)	44.44%	44.44%
LWAVVLLCF AVVLLCFVV	9	0.00% (0/1)	44.44%	44.44%
PGENITLSI DSRLQY	15	0.00% (0/1)	40.00%	40.00%
GTMAYGYG LNATILQ	15	0.00% (0/1)	33.33%	33.33%
VLLCFVV LIARAFYV	15	0.00% (0/1)	46.67%	46.67%
AQIIGLTN SEGQGIE	15	0.00% (0/1)	33.33%	33.33%

TABLE 4.18: Total Population Coverage of Designed Vaccine: Pakistan

MHC Class	Population Coverage	Average Hit	PC50
MHC-I and MHC-II (combined)	96.84%	11.64	8.42

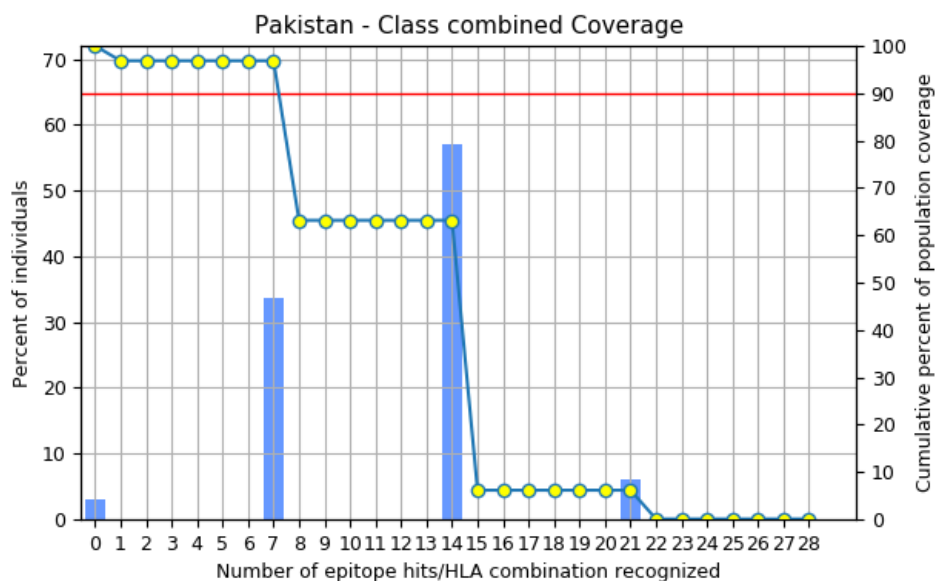


FIGURE 4.1: Graphical Representation of Population Coverage by MHC restricted epitopes in World.

TABLE 4.19: Total Population Coverage of Designed Vaccine: World

MHC Class	Population Coverage	Average Hit	PC50
MHC-I and MHC-II (Combined)	99.73%	20.54	15.51

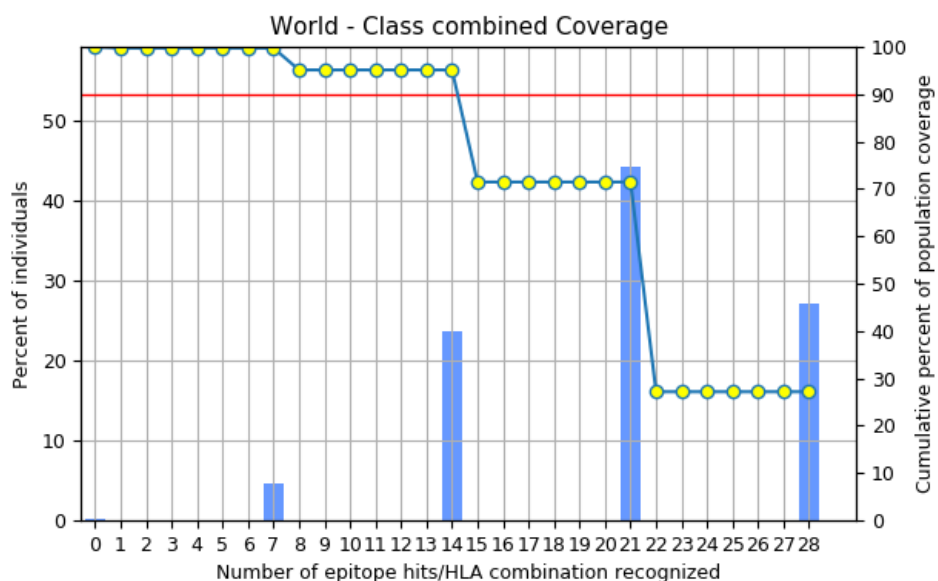


FIGURE 4.2: Graphical Representation of Population Coverage by MHC restricted epitopes in World.

4.7 Multivalent Vaccine Design Construction

The sequence of final vaccine composed of total 410 amino acids with overall three B-cell epitopes come after seven MHC-II (HTL) epitopes and seven MHC-I (CTL) epitopes [23]. The epitopes were detached by their defined linkers. Cholera enterotoxin subunit B was attached as an adjuvant at the N-terminus of vaccine construct for increasing the immunogenicity and at C-terminus of the vaccine, six histidine tag were added.

EAAAK linker is used as a rigid linker in the multivalent vaccine [25]. One of the main advantages of the use of linkers is that it maintains its specific functions from epitopes by preserving the structured separation of its functional domains. B-cell epitopes were linked by KK linker (Lysine linker). The KK linker was selected mainly because of three reasons:

1. Immunogenic responses was enhanced by KK linkers.
2. In the MHC-II restricted antigen presentation; lysosomal proteases like the Cathepsin B are involved in the peptide presentation on the surface of the cell through antigen processing of the peptides. These lysosomal proteases target KK linker.
3. The junctional immunogenicity was reducing through avoiding the antibodies induction.

For the attachment of the HLA-II epitopes with each other, GPGPG linker that is a universal spacer was used [24]. Study shows that performed GPGPG linker has ability of inducing TH lymphocyte (HTL) and has capability to split the epitopes through junctional immunogenicity that restores the immunogenicity of the individual epitopes. Subsequently, AAY (Ala-Ala-Tyr) linker was used to link MHC-I restricted epitopes. In mammalian cells it work as a cleavage site for the proteasomes so AAY linker was used to maintain essential junctional immunogenicity between the HLA-I epitope. Immunogenicity of the multi-epitope vaccine was also enhanced by AAY linker [24]. The sequence and order of epitopes in multivalent vaccine construct is shown below fig.

N^o-
 MIKLLKFGVFFTVLLSSAYAHGTPQNITDLCAEYHNTQIYTLNDKIFS YTESLAGKREMAII
 TFKNGAIFQVEVPGSQHIDSQKKAIERMKDTRLRIAYL TEAKVEKLCVWNNKTPHAIAAIS
 MANEAAKLWTPNGLTKGNENNAPKKFDMWRFYLLWAVVLLCCKEPEGENTLSIDSRL
 QYGPGPGAPFVFKPESTPAPKLGPGPGPFVFKPESTPAPKLDGPGPGFVFKPESTPAPKLD
 MGPGGPGGENITLSIDSRLQYGPGGGTMAYGYGLNATILQGPGGPVLLCFVVL IARAFY
 VGPGGPAQIIGLTNSEGQGIEAAYFYPPFPDM AAYREPFYPPFAAYVQVDSVTEEAA
 YWAVVLLCFVAAYVLLCFVVLIAAYLWAVVLLCF AAYAVVLLCFVHHHHHH
 -C^o

FIGURE 4.3: Order of peptides in multivalent vaccine construct. Sequence of the epitopes and linkers was marked by different colors (Aqua colored residues: Cholera enterotoxin subunit B, yellow residues: B-cell epitopes; Pink residues: MHC-II restricted epitopes (HTL); Turquoise residues: MHC-I restricted epitopes (CTL), blue, green and orange residues: linkers.

<u>Vaccine I</u> Vaccine with adjuvant and EAAK linker	<u>Vaccine II</u> Vaccine without adjuvant and EAAK linker
N ^o - MIKLLKFGVFFTVLLSSAYAHGTPQNIT DLCAEYHNTQIYTLNDKIFS YTESLAG KREMAIITFKNGAIFQVEVPGSQHIDS QKKAIERMKDTRLRIAYL TEAKVEKLC VWNNKTPHAIAAISMANEAAKLWTP NGLTKGNENNAPKKFDMWRFYLLW AVVLLCCKEPEGENTLSIDSRLQYGP PGAPFVFKPESTPAPKLGPGPGPFVFK PESTPAPKLDGPGPGFVFKPESTPAPK LDMGPGGPGGENITLSIDSRLQYGP GGTMAYGYGLNATILQGPGGPVLLC FVVL IARAFYVGPGGPAQIIGLTNSEG QGIEAAYFYPPFPDM AAYREPFYPP FFAAYVQVDSVTEE AAYWAVVLLCF VAAYVLLCFVVLIAAYLWAVVLLCF AAYAVVLLCFVHHHHHH -C ^o	N ^o - LWTPNGLTKGNENNAPKKFDMWRF YLLWAVVLLCCKEPEGENTLSIDSRL QYGPGPGAPFVFKPESTPAPKLGPG GPFVFKPESTPAPKLDGPGPGFVFKPE STPAPKLDGPGPGGENITLSIDSRL QYGPGPGTMAYGYGLNATILQGP PVLLCFVVL IARAFYVGPGGPAQII GLTNSEGQGIEAAYFYPPFPDM AAYR EPFYPPFAAYVQVDSVTEE AAYWA VVLLCFVAAYVLLCFVVLIAAYLWA VVLLCF AAYAVVLLCFVHHHHHH -C ^o

FIGURE 4.4: Graphical Presentation of Vaccine I and Vaccine II

4.7.1 Antigenicity and Allergenicity Prediction of Multivalent Vaccine Construct

VaxiJen v2.0 predicts the antigenicity of the vaccine construct at threshold of 0.5) was used. The antigenic score of the protein sequence of vaccine construct with adjuvant and EAAK linker generated by Vaxijen v2.0 was 0.8562 whereas the vaccine without adjuvant and EAAK linker was 1.0571. This result shows that both the vaccine proteins were antigenic. The allergenicity of protein vaccine was predicted by using AllerTOP v2.0 server. It predicts the allergenic potential of vaccine. The result obtained from this tool shows that the vaccine construct with adjuvant and EAAK linker and without adjuvant and EAAK linker were non-allergen.

TABLE 4.20: Antigenicity and allergenicity of multivalent vaccine

Sr.#	Property	Vaccine with adjuvant and EAAK linker	Vaccine without adjuvant and EAAK linker
1	Antigenicity	0.8562 (Probable ANTIGEN)	1.0571
2	Allergenicity	Non-allergen	Non-allergen

4.7.2 Physicochemical Properties and Solubility Determination of Primary Structure of Vaccine Construct

ProtParam tool was used to evaluate the physicochemical properties of the vaccine construct, including its molecular weight, theoretical isoelectric point value (PI), instability index, aliphatic index and GRAVY (grand average of hydrophaticity). The result of physiochemical properties obtained from ProtParam shows the molecular weight of protein with adjuvant and EAAK linker was 44750.94kDa and of protein without adjuvant and EAAK linker was 30412.33kDa. The theoretical pI value shows the nature of the protein. If the theoretical pI value of the protein is below 7, the nature of protein will be negative and if the theoretical pI value

of the protein is above 7, the nature of protein will be positive. The theoretical pI value of vaccine construct with adjuvant and EAAK linker was 7.12, so it is positive whereas theoretical pI value of vaccine construct without adjuvant and EAAK linker was 6.07, so it is negative. The instability index is used to determine whether the protein will be stable in a test tube. If the stability index of protein is less than 40, the protein will be stable in test tube and if the stability index of protein is above 40, the protein will not be stable. The instability index of vaccine protein with adjuvant and EAAK linker was 36.80, so it is stable and of protein without adjuvant and EAAK linker was 38.88, so it is also stable. The aliphatic index of protein with adjuvant and EAAK linker was 88.78 and of protein without adjuvant and EAAK linker was 88.55. The GRAVY value of a protein determines its hydrophobicity or hydrophilicity.

If the GRAVY score of a protein is below 0, it is considered as hydrophilic and if the GRAVY score of a protein is above 0, it is considered as hydrophobic. The GRAVY score of vaccine protein with adjuvant and EAAK linker was 0.135 of protein and GRAVY score of protein without adjuvant and EAAK linker was 0.246. Both the scores were above 0, so both the vaccine proteins was considered as hydrophobic. The estimated half life of vaccine protein with adjuvant and EAAK linker is 30 hours in mammalian reticulocytes if it tested in vitro whereas, if tested in vivo it is greater than 20 hours in yeast and greater than 10 hours in *Escherichia coli*. The estimated half life of vaccine protein without adjuvant and EAAK linker is 5.5 hours in mammalian reticulocytes if it is tested in vitro whereas, if tested in vivo it is 3 min in yeast and 2 min in *Escherichia coli*.

The solubility of vaccine construct was obtained from SOLUPROT v1.0. This predicts the expression of vaccine in *E.coli*. For the high expression of protein, the solubility should be above 0.5. If the solubility of protein is below 0.5, the protein will be insoluble. The protein solubility value of multivalent vaccine construct of protein with adjuvant and EAAK linker is 0.837 whereas, the protein solubility value of multivalent vaccine construct of protein without adjuvant and EAAK linker is 0.289. The result indicates that the protein sequence of vaccine with adjuvant and EAAK linker is soluble while that without adjuvant and EAAK linker is insoluble.

TABLE 4.21: Physio-chemical properties and solubility of multivalent vaccine

Sr.#	Property	Vaccine with adjuvant and EAAK linker	Vaccine without adjuvant and EAAK linker
1	Molecular weight	44750.94kDa	30412.33
2	Number of amino acids	410	282
3	pI value	7.12	6.07
4	Instability index	36.80	38.88
5	Aliphatic index	88.78	88.55
6	Gravy	0.135	0.246
7	Half life estimation	In vitro: 30 hours (mammalian reticuloocytes). In vivo: > 20 hours (<i>yeast</i>) and >10 hours (<i>E.coli</i>)	In vitro: 5.5 hours (mammalian reticuloocytes). In vivo: 3 min (<i>yeast</i>) and 2 min (<i>E. coli</i>)
8	Protein solubility	0.837	0.289

4.8 Structure Prediction, Refinement and Validation of Multi-epitope Vaccine

4.8.1 Secondary Structure Prediction

The vaccine secondary structure was determined by PSI-blast based secondary structure PREDiction (PSIPRED tool). Yellow color indicated the strand, pink color indicates the alpha helix, and grey color indicates coil. Vaccine protein secondary structure shows that it is stable.



FIGURE 4.5: Secondary Structure of Vaccine I

4.8.2 Tertiary Structure Prediction

The 3D structure of multivalent vaccine construct is predicted using i-TASSER. This tool initiates top five final models of the multivalent vaccine construct. Confidence score (C-score) describe the quality of the every predicted models. So, top C-score shows higher confidence in models that were generated. Besides C-score, iTASSER also calculated TM-score and RMSD value for all the five predicted

models of multivalent vaccine design construct, which predicts overall quality of model and helps in the choice of the model [27]. In this study, the C-score of first model tertiary structure higher than all other four models. The C-score for model 1 is -0.41. The estimated TM-score for model-1 is 0.66 and estimated RMSD is 8. So model-1 was selected.

4.8.3 Refinement of Vaccine 3D Structure

The predicted tertiary structure model from iTASSER was refined by using GalaxyRefine and trRosetta. The GalaxyRefine tool based on refinement method when used for refining the selected model-1 generated through iTASSER prediction server by improving both global and local structure quality on average. GalaxyRefine provides five additional models generated by relaxation simulations of the iTASSER model. Model 1 was selected.

4.8.4 Validation of Vaccine 3D Structure

The trRosetta tool is also used to refine the tertiary structure predicted model from iTASSER. Transform-restrained Rosetta (trRosetta) server is a web-based platform used for speedy and precise protein structure prediction. It generates five models of the iTASSER model. For the multivalent vaccine designed protein, if the predicted structure models by trRosetta are in low confidence (e.g., having estimated TM-score <0.5), it is possible that this design is not foldable, and wet-lab experiments are not required. The estimated TM-score for model-1 is 0.136 (<0.5). So model-1 was selected.

The obtained refined structures of vaccine protein from trRosetta and Galaxyweb were validated through Ramachandran plot using Ramachandran plot server. Results obtained from Ramachandran plot server for galaxyweb model shows that highly preferred observations were 96.707%, preferred observations were 3.293%, and questionable observations were 0.000%. So, model-1 of GalaxyWeb was selected for docking. The Errat value was also evaluated. The ERRAT value before refinement i.e., errat value of model-1 of i-TASSER was 90.452 and errat value of

after refinement i.e., errat value of model-1 of galaxyweb was 84.085. The Z-score of refined structure of (GalaxyWeb) predicted through molprobtity server was 2.05 ± 0.36

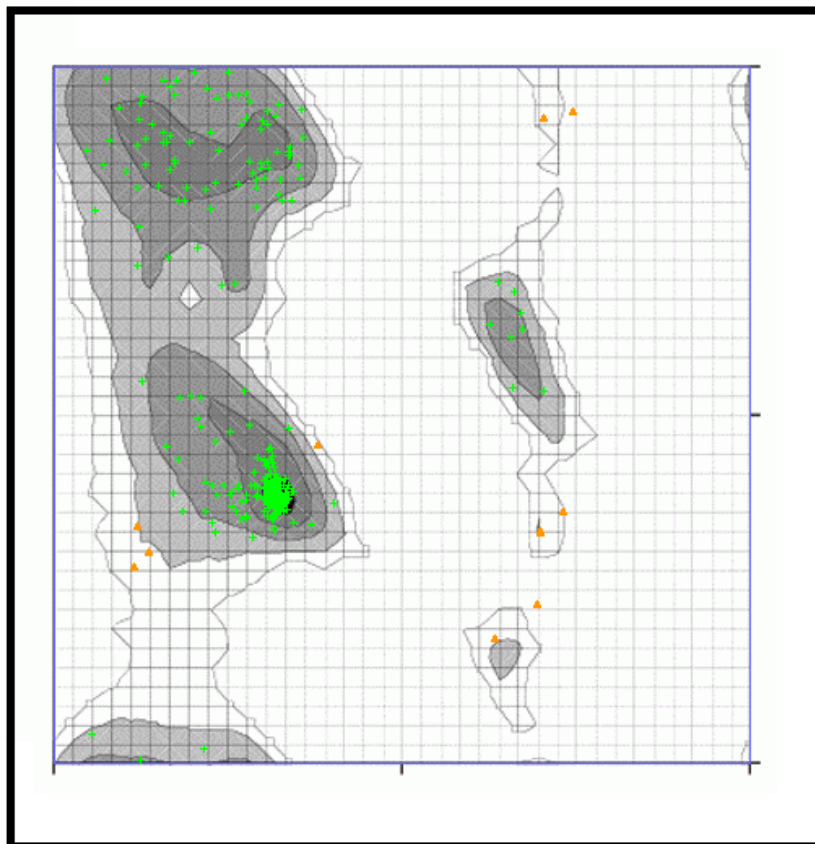


FIGURE 4.6: Ramachandran Plot of Refined 3D Structure (GalaxyWeb) of Vaccine I

Program: ERRAT2
File: model1 (i-tasser).pdb
Chain#:A
Overall quality factor**: 90.452

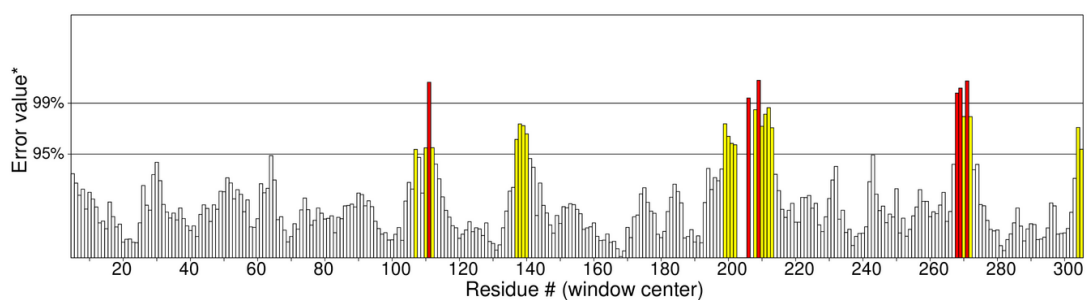


FIGURE 4.7: ERRAT Plot of 3D Structure (iTasser) of Vaccine I before refinement

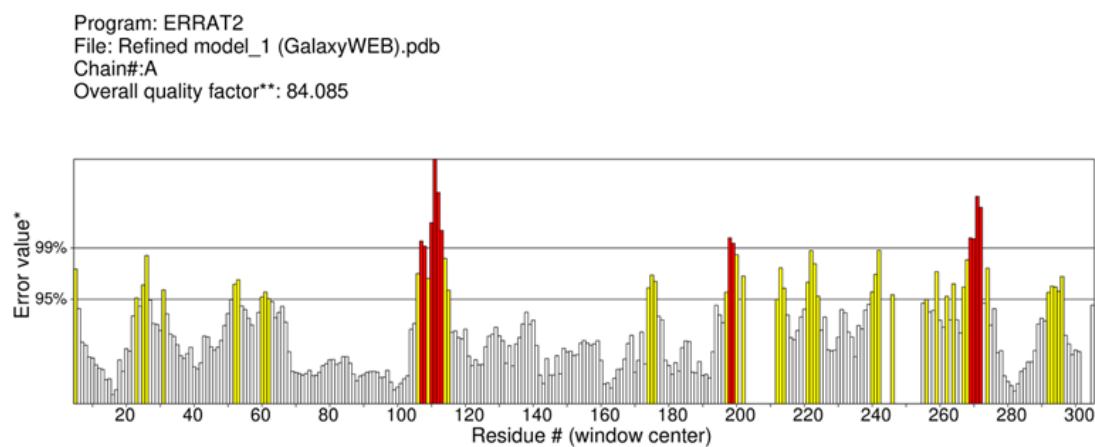


FIGURE 4.8: ERRAT Plot of 3D Structure (GalaxyWeb) of Vaccine I after refinement

4.8.5 Proteasomal Cleavage Sites Prediction of MHC-ligands

NetChop is a tool to predict cleavage sites of the human proteasome. The NetChop algorithm uses a neural network trained on human proteasome data. The NetChop-3.1 predicts that in the vaccine protein sequence with 410 amino acids polypeptide, the number of cleavage sites was 132.

4.9 Molecular Docking of Vaccine Protein with TLR-2 and its Structural Stability

The binding affinity between ligand and receptor molecules is assessed by molecular docking which is an insilico technique. In this study, Human toll-like receptor 2 (TLR2) was used to predict the binding affinity between the human TLRs and vaccine construct [5] [29].

The 3D structure of TLR2 retrieved from the Protein Data Bank (PDB) (PDB ID: 2z7x). The docking was done by use of ClusPro. In ClusPro the vaccine pdb structure retrieved from GalaxyWeb was uploaded in ligand and 2z7x (TLR2) receptor was uploaded in receptor option and then selected docked. The ClusPro generates 10 models of docking of vaccine with TLR2. The model 0 was selected

on the basis of energy level. The interactions between the ligand (vaccine) and receptor (2z7x) was determined through PDBePISA. The result from tool indicates vaccine I with adjuvant shows better interactions.

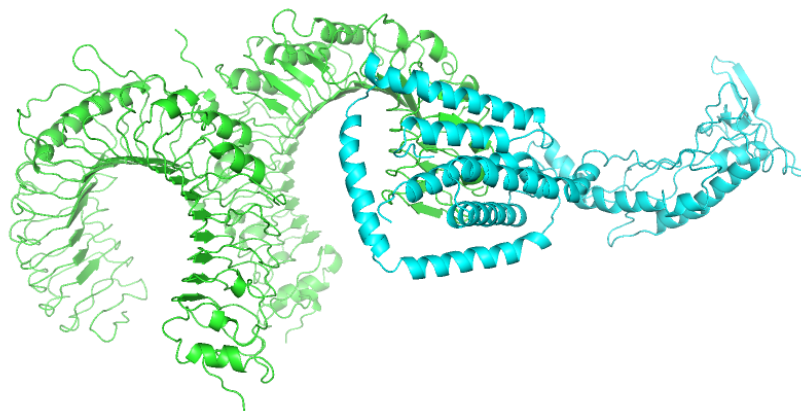


FIGURE 4.9: Protein-Protein interaction of Vaccine I (Ligand: Blue) with TLR2 (Receptor: Green) via ClusPro

TABLE 4.22: Predicted hydrogen bonding of docked vaccine I with TLR2 via PDBePISA

Sr.#	Structure 1	Dist.[Å]	Structure 2
1	A: TYR 376 [HH]	1.83	B: SER 309 [O]
2	A: LYS 347 [HZ3]	1.70	B: THR 361 [OG1]
3	A: LYS 347 [HZ1]	1.78	B: THR 363 [OG1]
4	A: GLN 396 [HE22]	2.46	B: SER 409 [OG]
5	A: ASN 345 [OD1]	1.67	B: LYS 385 [HZ2]
6	A: GLU 369 [OE1]	1.81	B: LYS 385 [HZ3]
7	A: GLU 374 [OE1]	2.05	B: ARG 337 [HH21]
8	A: GLU 375 [OE2]	1.79	B: ARG 337 [HH22]
9	A: GLU 375 [OE2]	2.12	B: ARG 337 [HE]

TABLE 4.23: Predicted salt bridges of docked vaccine I with TLR2 via PDBePISA

Sr. #	Structure 1	Dist.[Å]	Structure 2
1	A: GLU 369 [OE1]	2.69	B: LYS 387 [NZ]

2	A: GLU 374 [OE1]	2.78	B: ARG 337 [NH2]
3	A: GLU 375 [OE1]	3.26	B: ARG 337 [NH2]
4	A: GLU 375 [OE2]	2.69	B: ARG 337 [NH2]
5	A: GLU 375 [OE2]	2.93	B: ARG 337 [NE]

4.10 Immune-simulation

C-ImmSim server predicts the immune-simulation of the multivalent vaccine construct. The immunogenic reaction of the multivalent vaccine design construct was stimulated through C-ImmSim tool was used. The immune response of vaccine was predicted in both ways: with adjuvant and without adjuvant and EAAK linker. The result shows vaccine with adjuvant show better immunogenic response [22].

C-IMMSIM, an online web-server was used for immune simulation of vaccine construct. The C-ImmSim online server allows the user to define the antigen to be injected as a list of UniProt accession numbers, or PDB primary identifiers, or, as a multi-protein FASTA text. The haplotype is defined by drop-down menus. Other parameters are the simulation time and the volume.

The most important feature is the possibility to specify the antigen in terms of its constituting proteins. The user can submit the antigen protein sequence and specify the schedule of its injection as well as the host haplotype. Finally, other input fields allow the user to specify the random initial repertoire, the simulation duration and the simulated volume.

It simulates the response of host immune system to the multivalent vaccine design construct. The server is based on modeling approach and estimated the effect caused by foreign particle or antigen on the immune system using PSSM method (A PSSM is an abstraction of a multiple alignment of related sequences) [24]. C-IMMSIM calculated the production of cytokines, interferon and antibodies after vaccine injection.

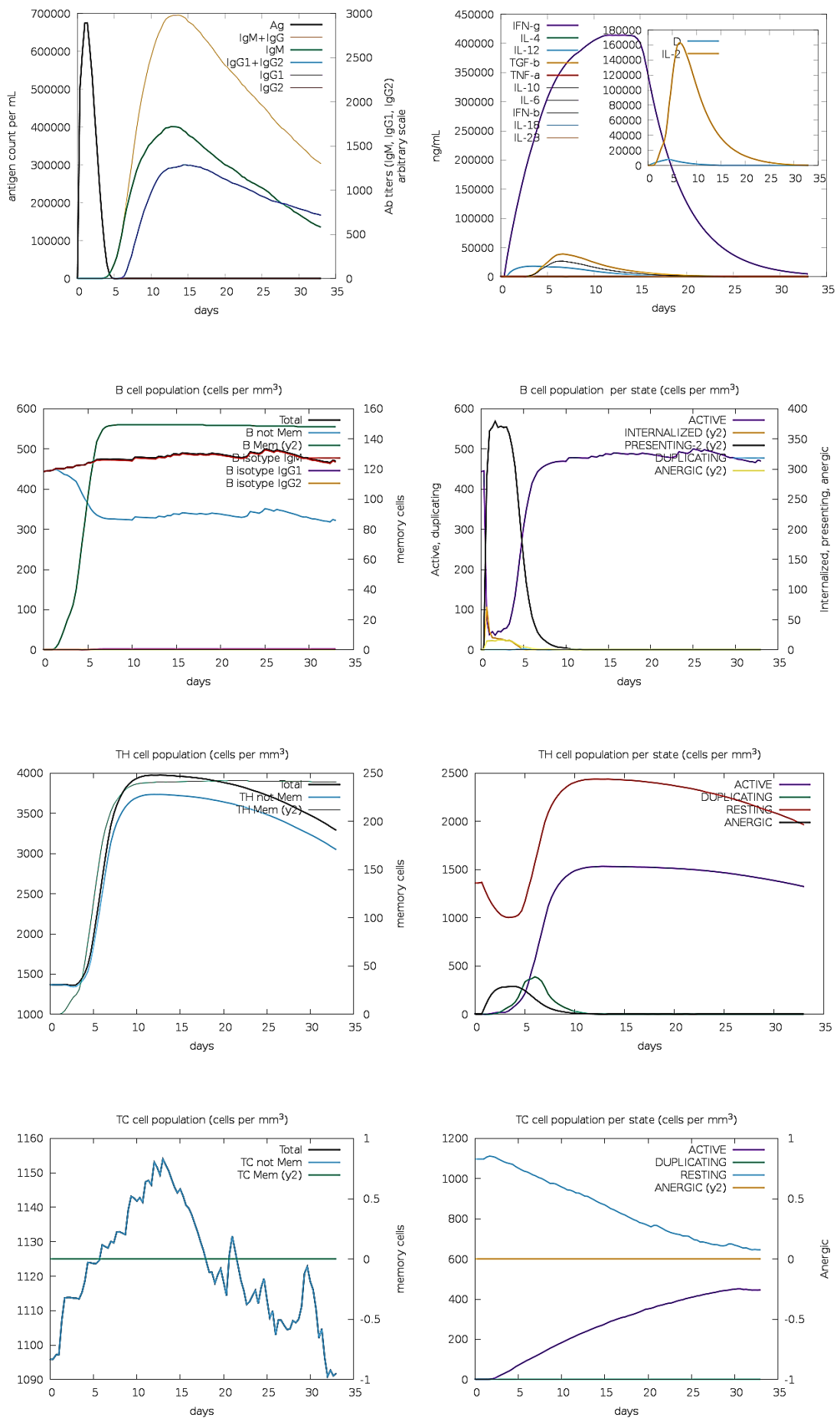


FIGURE 4.10: In-Silico Immune Simulation of Vaccine with Adjuvant (Vaccine I) predicted through C-IMMSIM tool following Injection of Vaccine; (a) Antigen and immunoglobulins: antibodies production by antigen and their sub-division. (b) Cytokines: rise in the concentration of cytokines and interleukins. D shows the danger signal along with higher production of IL-2 (growth factor) (c) Total B-cell population count (d) Total B-cell population count per state (e) Total T-helper cell (CD4) count (f) Total population of TH cell per state (g) Total population count of T-cytotoxic cell count (h) Total TC cell population count per state.

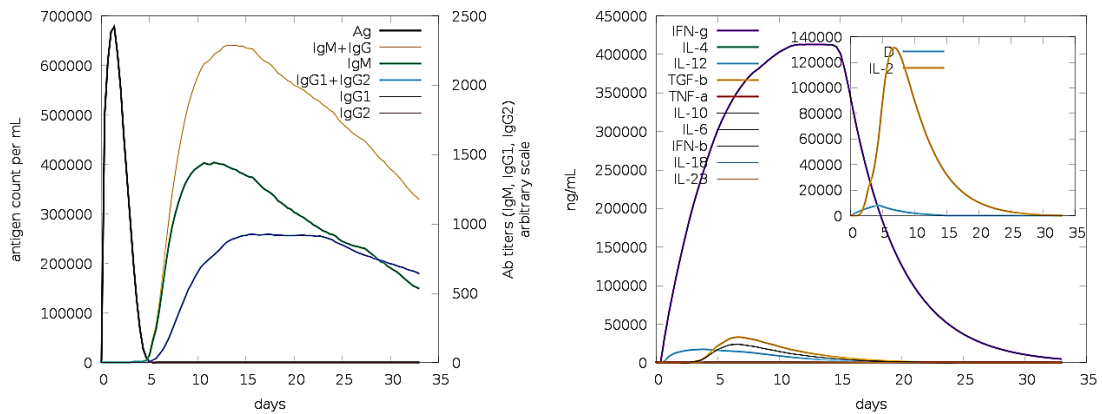


FIGURE 4.11: In-silico Immune Simulation of Vaccine without Adjuvant (Vaccine II) predicted through C-IMMSIM Tool following Injection of Vaccine; (a) Antigen and immunoglobulins: antibodies production by antigen and their sub-division. (b) Cytokines: rise in the concentration of cytokines and interleukins. D shows the danger signal along with high production of IL-2 (growth factor)

4.11 Gene Cloning

4.11.1 Sequence Translation of Vaccine Protein and Codon Adaptation

For the expression analysis of vaccine protein, the protein sequence of vaccine was translated. EMBOSS Backtranseq webserver was used to translate the protein sequence of the vaccine with adjuvant and EAAK linker. The result from the server shows that the protein (410 amino acids) was translated to 1230 nucleotides. After the translation of vaccine protein, codon adaptation was performed to improve DNA sequence of vaccine. The codons used in the vaccine structure were matched

to codons of *E.coli* (K12) using codon adaptation tool JCat. The improved DNA sequence shows the GC content was 53.41% and the CAI score was 0.956%.



FIGURE 4.12: Codon adaptation of Improved DNA.

4.11.2 In-Silico Gene Cloning with the Expression System *E.coli* K12

The length of improved codon sequence obtained from JCat was 1230bp with improved GC content and CAI score. This improved codon has been inserted into multiple cloning sites (MCS) of *E.coli* vector. The pET-28a(+) vector of *E.coli* was used that is 5369bp long. The improved vaccine codon was inserted between the restriction sites BssSI (3665) and PciI (3224). The final cloned product obtained from cloning is 5637bp long (5.637kbp).

This process was the in-silico gene cloning of codon optimized vaccine with the Expression system *E.coli* K12; pET-28a(+). The vaccine clone was obtained by integrating the vaccine fragment into vector pET-28a(+) of the *E.coli* vector between the restriction sites BssSI (3665) and PciI (3224).

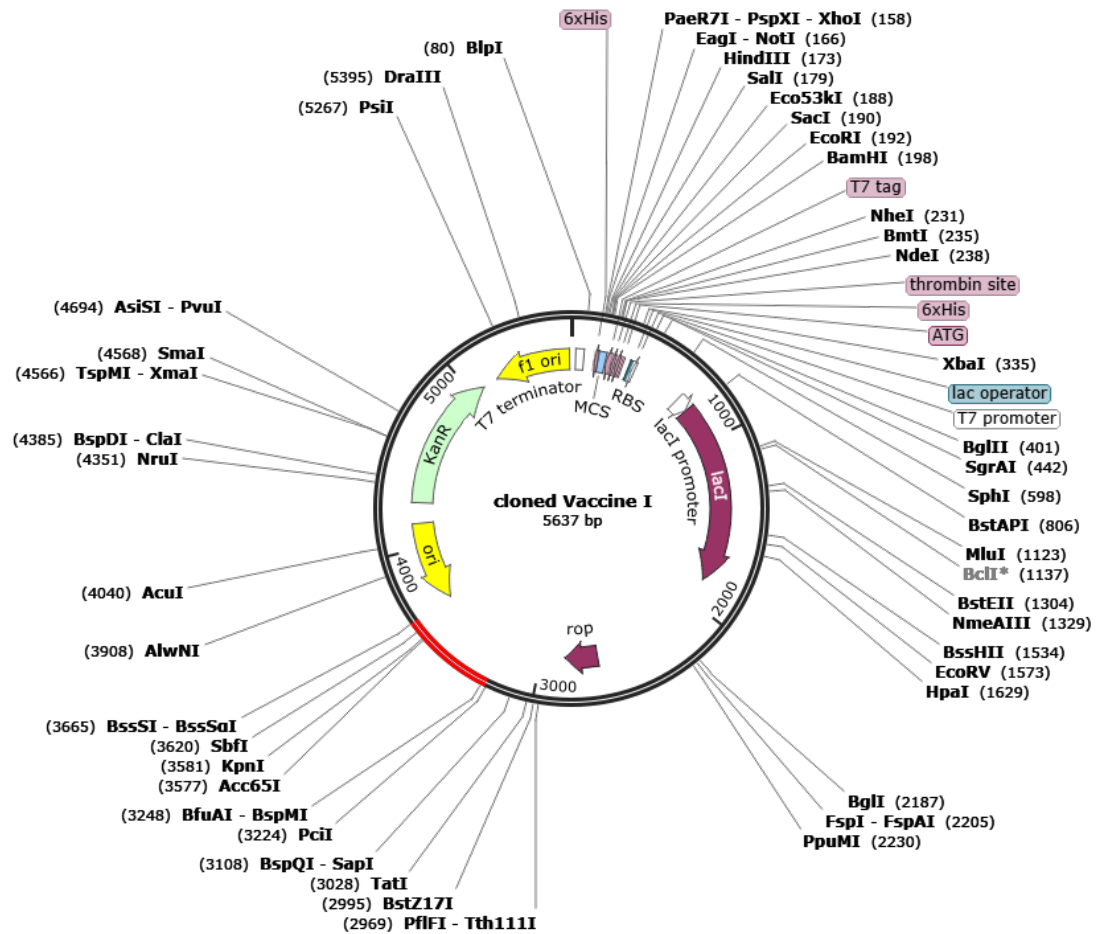


FIGURE 4.13: In-Silico Gene Cloning of codon optimized vaccine with the Expression system E.coli K12; pET-28 (+). Vaccine clone was obtained by integrating the vaccine fragment into vector pET-28a (+) of the E.coli vector between the restriction sites BssSI (3665) and PciI (3224). Plasmid is shown in black color and DNA sequence of vaccine is shown in red color.

Chapter 5

Conclusion and Future Prospects

Bacteria are small organisms with a single cell and found almost everywhere on Earth. Human body is consists of numerous bacteria. Many of them present in the body are not harmful, and few of them are beneficial. Only a small number of species cause disease.

Streptococci are gram positive and may be cocci or spherical in shape. Besides harmless species, the genus also comprises important pathogens of humans. The main specie are oral streptococci or viridans streptococci are found in the saliva and soft tissues of the oral cavity and play an essential role in the development of the micro-biome of mouth. Some species of oral streptococci also contribute to the development of plaque and can sometimes lead to infection. Viridans streptococci are α -hemolytic and play essential role in the formation of dental plaque, mainly *Streptococcus gordonii*, *Streptococcus sanguinis*, and *Streptococcus mitis*. They cause several diseases including infections, infectious endocarditis, and septicemia.

S. gordonii, is member of the group of viridians bacteria (α -hemolytic-sanguinis). It is Gram positive, mutualistic bacterium found in the human body, involving the oral cavity, upper respiratory tract, skin, and intestines. It is also opportunistic pathogen that can cause various diseases and infections including infective endocarditis and apical periodontitis. *S. gordonii* enters the inside of tooth i.e., root canal and vessels of blood, after which they interact with numerous immune (leukocytes) and non-immune cells of the body. Antibiotic treatment is available

against infection caused by *S. gordonii*. But antibiotic treatment is not long lasting and infection can be occurred again. Vaccine has long lasting impact on the organism to prevent infectious diseases. So by developing vaccine, one can develop immunity against *S. gordonii* and prevent infections and diseases caused by it like infective endocarditis (IE).

To develop the vaccine against infective endocarditis, surface proteins will be selected. The two surface proteins were selected according to choosen criteria i.e., antigenicity above 0.4, non-allergicity. The antigenic and non-allergen B cell and T cell epitopes selected from the proteins were used to construct multivalent vaccine with the help of linkers and adjuvant. The properties and structure of the protein was predicted. The predicted tertiary structure of the protein was validated and refined. The refined 3D structure of protein was docked with TLR2. The best docked structure was selected and interaction was predicted.

The future perspective of this study is to validate both proteins in wet lab as various drugs are determined in-silico and then validated by wet lab techniques are now a day's used as medicine.

Bibliography

- [1] “Bacteria,” Genome.gov. [Online]. Available: <https://www.genome.gov/genetics-glossary/Bacteria>. [Accessed: 09-Apr-2022].
- [2] K. J. Ryan and C. G. Ray, Sherris medical microbiology: An introduction to infectious diseases, 4th ed. McGraw-Hill Companies, 2004.
- [3] Colgate, “*Streptococcus gordonii*: Beware of these bacteria!,” Colgate.com, 16-Feb-2022. [Online]. Available: <https://www.colgate.com/en-us/oral-health/plaque-and-tartar/streptococcus-gordonii-bacteria>. [Accessed: 10-Apr-2022].
- [4] J. Abranches et al., “Biology of oral streptococci,” Microbiol. Spectr., vol. 6, no. 5, 2018.
- [5] O.-J. Park et al., “*Streptococcus gordonii*: Pathogenesis and host response to its cell wall components,” Microorganisms, vol. 8, no. 12, p. 1852, 2020.
- [6] Wikipedia contributors, “*Streptococcus gordonii*,” Wikipedia, The Free Encyclopedia, 14-Feb-2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Streptococcus_gordonii&oldid=1071772102.
- [7] W. Zheng, M. F. Tan, L. A. Old, I. C. Paterson, N. S. Jakubovics, and S. W. Choo, “Distinct Biological Potential of *Streptococcus gordonii* and *Streptococcus sanguinis* Revealed by Comparative Genome Analysis,” Sci. Rep., vol. 7, no. 1, p. 2949, 2017.
- [8] N. Mosailova, J. Truong, T. Dietrich, and J. Ashurst, “*Streptococcus gordonii*: A Rare Cause of Infective Endocarditis,” Case Rep. Infect. Dis., vol. 2019, pp. 1–2, 2019.

- [9] F. Garnier, G. Gerbaud, P. Courvalin, and M. Galimand, "Identification of clinically relevant viridans group streptococci to the species level by PCR," *J. Clin. Microbiol.*, vol. 35, no. 9, pp. 2337–2341, 1997.
- [10] J. O. Mundt, "The ecology of the streptococci," *Microb. Ecol.*, vol. 8, no. 4, pp. 355–369, 1982.
- [11] L. Chávez de Paz, G. Svensäter, G. Dahlén, and G. Bergenholtz, "Streptococci from root canals in teeth with apical periodontitis receiving endodontic treatment," *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.*, vol. 100, no. 2, pp. 232–241, 2005.
- [12] R. Rajani and J. L. Klein, "Infective endocarditis: A contemporary update," *Clin. Med.*, vol. 20, no. 1, pp. 31–35, 2020.
- [13] T. J. Cahill et al., "Challenges in infective endocarditis," *J. Am. Coll. Cardiol.*, vol. 69, no. 3, pp. 325–344, 2017.
- [14] S. Arshad, S. Awan, S. S. Bokhari, and M. Tariq, "Clinical predictors of mortality in hospitalized patients with infective endocarditis at a tertiary care center in Pakistan," *J. Pak. Med. Assoc.*, vol. 65, no. 1, pp. 3–8, 2015.
- [15] M. Tariq, M. Alam, G. Munir, M. A. Khan, and R. A. Smego Jr, "Infective endocarditis: a five-year experience at a tertiary care hospital in Pakistan," *Int. J. Infect. Dis.*, vol. 8, no. 3, pp. 163–170, 2004.
- [16] A. J. Pollard and E. M. Bijker, "A guide to vaccinology: from basic principles to new developments," *Nat. Rev. Immunol.*, vol. 21, no. 2, pp. 83–100, 2021.
- [17] R. Moxon, P. A. Reche, and R. Rappuoli, "Editorial: Reverse vaccinology," *Front. Immunol.*, vol. 10, p. 2776, 2019.
- [18] O. A. Dellagostin et al., "Reverse vaccinology: An approach for identifying leptospiral vaccine candidates," *Int. J. Mol. Sci.*, vol. 18, no. 1, p. 158, 2017.
- [19] M. Dalsass, A. Brozzi, D. Medini, and R. Rappuoli, "Comparison of open-source reverse Vaccinology programs for bacterial vaccine antigen discovery," *Front. Immunol.*, vol. 10, p. 113, 2019.

- [20] A. Caputo, P.-E. Fournier, and D. Raoult, “Genome and pan-genome analysis to classify emerging bacteria,” *Biol. Direct*, vol. 14, no. 1, p. 5, 2019.
- [21] V. P. Richards et al., “Phylogenomics and the dynamic genome evolution of the genus streptococcus,” *Genome Biol. Evol.*, vol. 6, no. 4, pp. 741–753, 2014.
- [22] M. Khan et al., “Immunoinformatics approaches to explore *Helicobacter Pylori* proteome (Virulence Factors) to design B and T cell multi-epitope subunit vaccine,” *Sci. Rep.*, vol. 9, no. 1, p. 13321, 2019.
- [23] V. Priyadarshini, D. Pradhan, M. Munikumar, S. Swargam, A. Umamaheswari, and D. Rajasekhar, “Genome-based approaches to develop epitope - driven subunit vaccines against pathogens of infective endocarditis,” *J. Biomol. Struct. Dyn.*, vol. 32, no. 6, pp. 876–889, 2014.
- [24] M. Sana, A. Javed, S. Babar Jamal, M. Junaid, and M. Faheem, “Development of multivalent vaccine targeting M segment of Crimean Congo Hemorrhagic Fever Virus (CCHFV) using immunoinformatic approaches,” *Saudi J. Biol. Sci.*, vol. 29, no. 4, pp. 2372–2388, 2022.
- [25] S. Aslam et al., “Proteome based mapping and reverse vaccinology techniques to contrive multi-epitope based subunit vaccine (MEBSV) against *Streptococcus pyogenes*,” *Infect. Genet. Evol.*, vol. 100, no. 105259, p. 105259, 2022.
- [26] M. Munia, S. Mahmud, M. Mohasin, and K. M. K. Kibria, “In silico design of an epitope-based vaccine against choline binding protein A of *Streptococcus pneumoniae*,” *Inform. Med. Unlocked*, vol. 23, no. 100546, p. 100546, 2021.
- [27] R. K. Pandey, R. Ojha, V. S. Aathmanathan, M. Krishnan, and V. K. Prajapati, “Immunoinformatics approaches to design a novel multi-epitope subunit vaccine against HIV infection,” *Vaccine*, vol. 36, no. 17, pp. 2262–2272, 2018.
- [28] S. Pyasi, V. Sharma, K. Dipti, N. A. Jonniya, and D. Nayak, “Immunoinformatics approach to design multi-Epitope- subunit vaccine against bovine ephemeral fever disease,” *Vaccines (Basel)*, vol. 9, no. 8, 2021.

-
- [29] H. Y. Kim, J. E. Baik, K. B. Ahn, H. S. Seo, C.-H. Yun, and S. H. Han, “*Streptococcus gordonii* induces nitric oxide production through its lipoproteins stimulating Toll-like receptor 2 in murine macrophages,” *Mol. Immunol.*, vol. 82, pp. 75–83, 2017.
- [30] U. K. Adhikari, M. Tayebi, and M. M. Rahman, “Immunoinformatics approach for Epitope-based peptide vaccine design and active site prediction against polyprotein of emerging Oropouche virus,” *J. Immunol. Res.*, vol. 2018, pp. 1–22, 2018.