CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Application of Pan-Genomics Approach for *Clostridium botulinum* Therapeutics

by

Iqra Riasat

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

2021

Every challenging work needs self-efforts as well as the guidance of elders. I dedicate this thesis to my parents whose affection makes me able to get such success and to my teachers whose encouragement has always been my source of inspiration.

CERTIFICATE OF APPROVAL

Application of Pan-Genomics Approach for *Clostridium botulinum* Therapeutics

by

Iqra Riasat

(MBS193022)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Mazhar Qayyum	PMAS, AAU, Rawalpindi
(b)	Internal Examiner	Dr. Arshia Amin Butt	CUST, Islamabad
(c)	Supervisor	Dr. Syeda Marriam Bakhtiar	CUST, Islamabad

Dr. Syeda Marriam Bakhtiar
Thesis Supervisor
September, 2021

Dr. Sahar Fazal
Head
Dept. of Bioinfo. and Biosciences
September, 2021

Dr. M. Abdul Qadir
Dean
Faculty of Health and Life Sciences
September, 2021

Author's Declaration

I, **Iqra Riasat** hereby state that my MS thesis titled "**Application of Pan-Genomics Approach for *Clostridium botulinum* Therapeutics**" 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.

(Iqra Riasat)
Registration No: MBS193022

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Application of Pan-Genomics Approach for *Clostridium botulinum* Therapeutics**" 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.

(Iqra Riasat)
Registration No: MBS193022

List of Publications

It is certified that following publication(s) have been made out of the research work that has been carried out for this thesis:-

1. **Iqra Riasat**,Syeda Marriam Bakhtiar, Muhammad Faheem, Arun Kumar Jaiswal, Muhammad Naeem, Raees Khan, Asmat Ullah Khan et al. (2021). Application of pan genomics towards the druggability of Clostridium botulinum. Applied Nanosciences , 1-13.

2. **Iqra Riasat**,Saira Hafeez, Sehrish Imtiaz*, Naveed Iqbal Soomro*, Syeda Marriam Bakhtiar* (2021). Role of rs6265 (G196a) Polymorphism of BDNF Gene in Pathogenesis of Depression. 1st International Hybrid Mode Conference on Emerging Innovate Reseearch Trends in Biology.

3. **Iqra Riasat**, s. M. (2021). Elucidating the Genetic Factors Contributing Towards Susceptibility and Severity Of Covid 19. Current Chinese Science Journal.

4. Hajra Qayyum, Anum Munir, Syeda Maham Fayyaz, Ayesha Aftab, Hina Aslam Butt, Naveed Iqbal Soomro,Sobia Khurshid, Muhammad Qasim Khan, Saeed Iqbal Soomro, Yumna Saghir, **Iqra Riasat**, Mohammad Nadeem and Syeda Marriam Bakhtiar. 1. (2019). Single Cell Omics in Cardiovascular Diseaes. Book Single Cell Omics Vol II (Applications in Biomedicine and Agriculture) ISBN: 978-0-12-817532-3.

5. Saira Hafeez,Amna Noor, **Iqra Riasat**, Muhammad Qasim, Syeda Marriam Bakhtiar (2018) "etiology, genetics, clinical manifestations and possible treatments of acromegaly" precision medicine vol 3 number 1 page 29-37.

6. Muhammad Qasim Khan, Nosheen Afzal, **Iqra Riasat,** Saira Hafeez, Sehrish Imtiaz, Samreen, Syeda Marriam Bakhtiar (2019). Emerging Trends of Single Cell Omics in Neuroscience Research. 1st International Conference on Recent Updates in Biotechnology ICRUB 2019.

7. Hina Aslam Butt, Syeda Marriam Bakhtiar, Amna Noor, **Iqra Riasat** (2018). Genetic Heterogeneity in Pakistani Obese Families. 1st Annual Health Reaserch Conference, Islamabad.

(Iqra Riasat)

Registration No: MBS193022

Acknowledgement

I would like to thanks Allah Almighty, The Most Magnificent and Compassionate, indeed, all praises are due to Him and His Holy Prophet Muhammad (PBUH). They gave me the strength and aptitude to complete this target.

I want to acknowledge the efforts of my thesis supervisor Dr. Syeda Marriam Bakhtiar, Department of Biosciences and Bioinformaticss, whose encouragement, guidance and support helped us to complete our project.

I want to give my sincere gratitude to my co-supervisor Dr. Syed Babar Jamal Bacha, National University of Medical Sciences who supported, motivated and guided me throughout my research journey and without whom it would have been difficult for me to complete this study. I want to say thank you to Dr. Muhammad Faheem for his continous help during my research work. I want to acknowledge Dean of Faculty of Health and Life Sciences, Dr. Muhammad Abdul Qadir and head of Department of Bioinformatics and Biosciences, Dr. Sahar Fazal, for giving me the opportunity to persue MS with thesis and complete my research within time. I owe a great deal of appreciation and gratitude to all the Faculty members, Dr. Shaukat Iqbal Malik, Dr. Erum Dilshad and Dr. Arshia Amin Butt. A special thanks go to all my friends and seniors for their support, coordination and help from time to time.

In the end, I am gratefully wanted to acknowledge my parents for their countless contributions, all their support without which I was unable to do anything. I am out of words to explain my gratitude towards my parents, siblings for their love, care, encouragement and prayers that enlightened my whole life.

(Iqra Riasat)

Abstract

Clostridium botulinum is anaerobic spore forming bacteria that produces very potent neurotoxins with the potential to be used as bioweapon. In this study, pan genome approach was utilized to develop drug targets for 51 strains of *Clostridium botulinum*. From the total 8756 core proteins, there were 422 non-host homologous proteins from which we selected 9 essential proteins by applying 2 thresholds that are identity greater than 35 and e-value =0.001. From these 7 proteins were selected as drug targets and docked against 105 anti-bacterial compounds. After docking, against each protein top 10 compounds are selected based on the binding affinity and from these 10 compounds 3D structure of one docked compound is shown that either have highest binding affinity or number of residues interaction. 2 proteins were selected for epitope-based study and 2 vaccines were manually developed on the bases of T-cell epitopes of these proteins. Both these vaccines were docked against 4 HLAs that play important role in human immune system. In the Drug-targeting study, there are also some compounds that has shown highest binding energies with more than one protein. From the 105 anti-bacterial compounds, I got 39 compounds on the basis of their binding energies with all target proteins. In the epitope-based study, vaccine 2 shows highest interaction with all the HLAs. These results can be further validated by in vitro analysis and can proceed for clinical trials.

Contents

Author's Declaration iv

Plagiarism Undertaking v

List of Publications vi

Acknowledgement viii

Abstract ix

List of Figures xiii

List of Tables xvi

Abbreviations xviii

1 Introduction 1
- 1.1 Drug Targeting Aprroach . . . 3
- 1.2 Reverse Vaccinology Approach . . . 3
- 1.3 Aims and Objectives . . . 4

2 Literature Review 6
- 2.1 *Clostridium botulinum* . . . 6
- 2.2 Epidemiology of Botulism Syndromes . . . 7
  - 2.2.1 Foodborne Botulism . . . 8
  - 2.2.2 Botulism of Wounds . . . 9
  - 2.2.3 Infant Botulism . . . 9
  - 2.2.4 Botulism with Adult Bowel Toxaemia . . . 10
  - 2.2.5 Inhalational botulism . . . 10
  - 2.2.6 Botulism of Iatrogenic . . . 10
- 2.3 Botulinum Neurotoxin . . . 11
- 2.4 Animal Botulism . . . 12
  - 2.4.1 Group III Animal and *C. Botulism* . . . 12
  - 2.4.2 Group III Presence of *C. Botulism* in Birds & Cattle . . . 13
- 2.5 Toxin Structure . . . 14

2.5.1 Complex Toxins . . . 14
2.5.2 Mode of Action . . . 15
2.5.3 Internalization . . . 17
2.5.4 Neurotransmitter Release . . . 17
2.6 Pan-Genome Analysis of *Clostridium botulinum* . . . 18
2.7 B-Cell Epitope Mapping . . . 20
2.8 Virulome . . . 22

3 Materials And Methods **24**
3.1 Drug Targeting Techniques . . . 24
3.1.1 Genome Selection . . . 24
3.1.2 Identification of Core Genome . . . 24
3.1.3 Non-homologous Protein Identification . . . 25
3.1.4 Target Identification . . . 25
3.1.5 Catalytic Pocket Detection . . . 26
3.1.6 Molecular Docking . . . 26
3.1.7 Docking Validation . . . 26
3.2 Epitope-Based Vaccine Identification . . . 26
3.2.1 Data Retrieval and Structural Analysis . . . 27
3.2.2 B-cell Epitope Prediction . . . 27
3.2.3 T-cell Epitope Prediction . . . 28
3.2.4 Imperative Features Profiling of Selected T Cell Epitopes . . . 28
3.2.5 Epitope Conservation Analysis . . . 29
3.2.6 Multi-epitope Vaccine Design and Construction . . . 29
3.2.7 Physiochemical Analysis of Multi-epitope Vaccines . . . 29
3.2.8 Multi-epitope Vaccines 3D Structure Prediction . . . 30
3.2.9 Molecular Docking . . . 30
3.3 Methodology Overview . . . 30

4 Results and Discussion **32**
4.1 Drug Targeting Analysis . . . 32
4.2 Molecular Docking . . . 33
4.2.1 ATP Synthase Subunit C . . . 34
4.2.2 Chromosomal Replication Initiator Protein DnaA . . . 36
4.2.3 Thil Domain-containing Protein . . . 38
4.2.4 Putative guanosine 3',5' -bis- pyrophosphate (PpGpp) . . . 41
4.2.5 Putative Metal Dependent Phosphohydrolase . . . 43
4.2.6 Putative Metal Dependent Phosphohydrolase . . . 45
4.2.7 UDP-N-acetylmuramoyl-tripeptide-D-alanyl-D- alanine Ligase 48
4.2.8 Chemotaxis . . . 50
4.3 Epitope-Based Vaccine Identification . . . 52
4.3.1 Epitope Target Identification . . . 52
4.3.2 Structure Analysis . . . 52
4.3.3 T-cell Epitope Recognition . . . 63

4.3.4 Imperative Features Profiling of Selected T Cell Epitopes . . 64
4.3.5 Epitope Conservation Analysis 65
4.3.6 Multi-epitope Vaccine Design and Construction 68
4.3.7 Physiochemical Properties of Multi-epitope Vaccines 69
4.3.8 Molecular Docking 69

5 Conclusion and Future Prospects **74**

Bibliography **76**

Appendices **91**

List of Figures

2.1 Stages of inhibition of *C. botulinum* in host organism 16
2.2 Possible Inhibition mechanisms of *C. botulinum*. 18

3.1 Methodology used for the drug-targeting and epitope-based study . 31

4.1 Docking of multiple ligands 34
4.2 Compound 2,3-dibromo-5,8-dihydroxy naphthalene-1,4-dione that shows Binding affinity with ATP synthase subunit C 36
4.3 Compound 3-methyl-1-(1,4,5,8-tetramethoxynaphthalen-2-yl) but -3-en-1- ol that shows Binding affinity with Chromosomal replication initiator protein DnaA. 38
4.4 Compound (E)-4-(4-bromo-2, 5-dimethoxyphenyl) -4-oxobut-2-enoic acid that shows residue interaction with Thil domain-containing protein. 39
4.5 Compound 2 - (2,5-dimethoxy-4 -methylbenzoyl) -3- methoxybenzoic acid that shows binding energy along with residue interaction with Putative guanosine 3',5'-bis-pyrophosphate (PpGpp) synthesis / degradation protein. 43
4.6 Compound 2- (2,5-dimethoxy-4-methylbenzoyl) -3- methoxybenzoic acid that shows binding energy along with residue interaction with Putative guanosine 3',5'-bis-pyrophosphate (PpGpp) synthesis/degradation protein. 45
4.7 Compound 2-(2, 5 -dimethoxy-4 - methylbenzoyl) -3 -methoxybenzoic acid that shows number of residue interaction with Putative metal dependent phosphohydrolase protein. 47
4.8 Compound 1,4,5,8-tetramethoxy-2-naphthoic acid that shows highest Binding affinity with UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D-alanine ligase. 48
4.9 Compound 2, 3-dibromo-5, 8-dihydroxynaphthalene-1, 4-dione that shows affinity with chemotaxis protein. 52
4.10 PSPIRED analysis of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) 56
4.11 PSPIRED analysis of Preprotein translocase, SecG subunit B-cell epitope recognition 57

4.12 Preprotein translocase, SecG subunit a Bepipred Linear Epitope prediction results, b flexibility analyses using Karplus and Schulz flexibility scale, c beta turns analyses in structural polyprotein using Chou and Fasman beta-turn prediction, d surface accessibility analyses using Emini surface accessibility scale, e prediction of antigenic determinants using Kolaskar and Tongaonkar antigenicity scale, f hydrophilicity prediction using Parker hydrophilicity 59
4.13 UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) a Bepipred Linear Epitope prediction results, b flexibility analyses using Karplus and Schulz flexibility scale, c beta turns analyses in structural polyprotein using Chou and Fasman beta-turn prediction, d surface accessibility analyses using Emini surface accessibility scale, e prediction of antigenic determinants using Kolaskar and Tongaonkar antigenicity scale, f hydrophilicity prediction using Parker hydrophilicity 60
4.14 Site of discontinuous epitopes predicted through Disctope on the 3D structure of **(A)** Preprotein translocase, SecG subunit and **(B)** UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) and presented by sticks through Pymole 61
4.15 Site of linear epitopes predicted through PepSurf on the 3D structure of **(A)** Preprotein translocase, SecG subunit and **(B)** UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) presented by balls 62
4.16 Multiple sequence alignment of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN) 67
4.17 Multiple sequence alignment of pre-protein translocase, SecG subunit 68
4.18 Graphical representation of VacI and VacII. 68
4.19 Protein-Protein interaction of 1BOR with Vac-1 (A) complex of Vac-I interacting with IBOR (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins. 70
4.20 Protein-Protein interaction of 1BOR with Vac-II (A) complex of Vac-II interacting with IBOR (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins. 70
4.21 Protein-Protein interaction of 5IPF with Vac-I (A) complex of Vac-I interacting with 5IPF (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins. 71
4.22 Protein-Protein interaction of 5IPF with Vac-II (A) complex of Vac-II interacting with 5IPF (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins. 71

4.23 Protein-Protein interaction of 6AT5.1.A with Vac-I (A) complex of Vac-I interacting with 6AT5.1.A (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins. 72
4.24 Protein-Protein interaction of 6AT5.1.A with Vac-II (A) complex of Vac-II interacting with 6AT5.1.A (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins. . 72
4.25 Protein-Protein interaction of 3BOR with Vac-I (A) complex of Vac-I interacting with 3BOR (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins. 73
4.26 Protein-Protein interaction of 6AT5.1.A with Vac-II (A) complex of Vac-II interacting with 6AT5.1.A (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins. . 73

List of Tables

2.1 Various types of Neurotoxins produced by different strains of C. botulinum 11

4.1 Validation scores of the selected proteins against Rampage and ERRAT 33
4.2 Compounds name, binding affinity and Residue interaction of top 10 compounds with ATP synthase subunit C 35
4.3 Compound's name, binding affinity and Residue interaction of top 10 compounds with Chromosomal replication initiator protein DnaA 37
4.4 Compounds name, binding affinity and Residue interaction of top 10 compounds with Thil domain-containing protein 39
4.5 Compound's name, binding affinity and Residue interaction of top 10 compounds with Putative guanosine 3',5'-bis-pyrophosphate (PpGpp) synthesis/degradation protein 41
4.6 Compound's name, binding affinity and Residue interaction of top 10 compounds with Putative guanosine 3',5'- bis-pyrophosphate (PpGpp) synthesis / degradation protein 44
4.7 Compounds name, binding affinity and Residue interaction of top 10 compounds with Putative metal-dependent phosphohydrolase protein 46
4.8 Compounds name, binding affinity and Residue interaction of top 10 compounds with UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D-alanine ligase 49
4.9 Compounds name, binding affinity and Residue interaction of top 10 compounds with Chemotaxis protein 50
4.10 Physiochemical parameters of both proteins computed via Protparam 53
4.11 Disulfide bonds along with distance, bond and bond score for UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) 54
4.12 B-cell epitopes present on the surface of Preprotein translocase, SecG subunit predicted via BCPRED along with their starting position and antigenicity scores. 59
4.13 B-cell epitopes present on the surface of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) predicted via BCPRED along with their starting position and antigenicity scores 59
4.14 Discontinuous epitopes of Preprotein translocase, SecG subunit predicted through DICOTOPE 2.0 server 60

4.15 Discontinuous epitopes of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) predicted through DICOTOPE 2.0 server 62
4.16 MHC class-I allele binding peptides of pre-protein translocase, SecG subunit predicted via Propred-1 with their antigenicity scores . . . 64
4.17 MHC class-I allele binding peptides of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN) predicted via Propred-1 with their antigenicity scores 64
4.18 Conservation of epitopes of pre-protein translocase, SecG subunit via IEDB epitope conservancy analysis tool 65
4.19 Conservation of epitopes of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN) via IEDB epitope conservancy analysis tool 66

5.1 Genomic Statistical features of selected strains of *Clostridium botulinum* 91
5.2 MHC class-II allele binding peptides of pre-protein translocase, SecG subunit predicted via Propred with their antigenicity scores 94
5.3 MHC class-II allele binding peptides of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN) predicted via Propred with their antigenicity scores 97
5.4 Biological and molecular characteristics of the essential proteins . . 101
5.5 Peptides of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN) with non-digesting enzymes, mutation position, toxicity, allergenicity, hydrophobicity, hydrophilicity, charge, and PI. 107
5.6 Peptides of pre-protein translocase, SecG subunit with non-digesting enzymes, mutation position, toxicity, allergenicity, hydrophobicity, hydrophilicity, charge, and PI 125

Abbreviations

A	Allergen
C. botulinum	*Clostridium botulinum*
Blast	Basic Local Alignment Search Tool
DEG	Database of Essential Genes
DIANNA	DIAminoacid Neural Network Application
Edgar	Efficient Database framework for comparative Genome Analyses using BLAST score Ratios
ERRAT	overall quality factor
GRAVY	Grand Average of Hydrophaticity
HLA	Human leukocyte antigen
KDA	Kilodalton
Kegg	Kyoto Encyclopedia of Genes and Genomes
MHC	Major Histocompatibility complex
MOE	Molecular Operating Environment
MW	Molecular Weight
NA	Non-allergen
NM	No mutation
NT	Non-toxin
PSIPRED	PSI-blast based secondary structure PREDiction
RAMPAGE	RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression
RMSD	Root Mean square devation
TMHMM	Transmembrane Helices; Hidden Markov Model
Uniprot	Universal Protein Resource

VAC	Vaccine
Vfdb	Virulence factor Database

Chapter 1

Introduction

Clostridium botulinum is a member of the firmicutes and is an an-anaerobic, heterogeneous, Gram-positive organism that produces spores [1]. This family produces several types of toxins that are collectively known as the botulinum neurotoxin. These toxins are the most potent known toxins towards man and have the ability to induce potentially fatal paralytic conditions in humans as well as in other animal species and the conditions are known as "botulism". The most common reported diseases in humans are food-borne botulism, infant botulism and wound botulinum. This disease has a very high fatality rate. As the botulinum toxins affect the nervous system due to which it is known as a neurotoxin.

In foodborne botulism descending, flaccid paralysis can take place that further causes the failure of the respiratory system. Fatigue, weakness and vertigo are considered as the early symptoms that will be followed by blurred vision, difficulty in swallowing and speaking with a dry mouth. Some other conditions of this disease also include diarrhea, vomiting, abdominal pain and constipation. As the condition advanced, it caused weakness in the neck and arms, which was followed by weakness in the respiratory muscles and lower body muscles. This sickness does not cause fever or loss of consciousness. It is the toxin, not the bacteria itself that causes these symptoms. Within the 12 to 36 hours after the exposure, symptoms start to appear. The minimum range is 4 hours while the maximum range is 8 days. Botulism incidence is very low but it has a high mortality rate in

case of inappropriate diagnosis and immediate treatment is not given. Immediate treatments include early administration of the antitoxin and intensive respiratory care. In 5 to 10% of the cases, this disease can be fatal [2].

For the considerable hazards of the botulinum neurotoxin-forming clostridia presented two biological aspects are responsible. Endospore's formation is their first ability that has high resistance. This property enables clostridia to survive some of the adverse treatments such as heating. Understanding the mechanism of spore heat resistance and the physiological processes that are involved in the germination and lag phase is the most important for the better control of the botulinum neurotoxin-forming clostridia. After this understanding, important progress is being made.

The second most important biological aspect is their ability to make highly potent neurotoxin. The most toxic poison known is the botulinum neurotoxin that can be fatal as little as 30-100ng [2]. Some of the aspects of this neurotoxin are being known by advanced information such as the crystallographic structure and regulatory mechanisms of the neurotoxin is still poorly understood. Intestinal toxaemia is infant botulism that affects children ages below 12 years. There are very rare chances of this disease to have occurred in adults. This disease occurs when the normal intestinal microbiota is suppressed by the competing bacteria [3]. In many countries, infant botulism is being reported but it is the commonest manifestation of the disease is in the United States [4].

Botulinum toxins produced by the *Clostridium botulinum* are of seven distinct forms are types A-G. From these seven forms A,B,E,F are four known forms which cause botulism occasionally, on the other hand in mammals, birds and fish it is caused by type [5] that has been categorized into I-IV physiological groups [6]. Through improperly processes food, botulinum toxins are ingested.

The bacteria or the spore that has been survived in t hat improperly processed food starts to grow and produce toxins within the body. Apart from food-borne intoxication, human botulism can also be caused by the inhalation of *clostridium botulinum* that can cause infection in infants, or some sort of wound infections [7].

1.1 Drug Targeting Aprroach

Computational methods have been developed [8, 9] in order to quickly identify new targets in the postgenomic period. The objectives for *Burkholeria pseudiomellei* [10], and *Cynobacterium Dipheria* [11], are comparative microbial and differential analyses of the genomes *Helicobacter pylori* [12], *Mycobacterium tuberculosis* [13], *Pseudomonas aeruginosa* [14], *Salmonella typhi* [15] and *Neisserisa gonorrhe* [16, 17]. In the drug discovery process, the identification of therapeutic targets is in earlier stages. When this process is completed, research on drug and target identification will step into new era [18, 19]. The availability of pathogen and host genome sequences on a genomic basis for any pathogens has facilitated the identification of drug targets [20]. For recognition of the drug targets where their role has been well defined, functional awareness of the individual protein is generally established. The expense of innovation in science has slowly and steadily increased in the last ten years in the pharmaceutical sector, but the time taken for the production of a new medication is the same as always, which is between 10 and 15 years [21].The pace of the ventures increases with the decrease in the expense that systems can now see entire microbes by in-silico methods. These advances enable scientists to do things that seem unlikely or difficult to deal with experimentally in the field of science. In the drug-discovery procedures, the evolution of the main molecular pathways of the diseases has been progressive, from classical ligand-based drug discoveries to structural and selective methods for drug design. With cell biology principles and an overall understanding of the microbes as a whole, new opportunities for identifying computational drug targets will be opened [22].

1.2 Reverse Vaccinology Approach

The manufacture of vaccines against clinically significant diseases is the most important application to immunology in public health. While traditional vaccine preparation strategies are focused on the reduced or inactivated whole virus or

partly purified virous protein, the intrinsic viral characteristics of such strategies include weak or null replication and antigenic hyper-variability of pathogens in vitro [23] have some disadvantages. To overcome these types of issues, several novel approaches have been developed and from these techniques, the most promising one is the epitope-based vaccine preparation. This technique uses several peptides such as peptides of beta cells, MHC complex peptides of class I and class II. Epitope based vaccine preparation technique possesses many advantages such as peptides can be easily produced in vitro that will reduce the production costs which will simplify the vaccine production procedures at a large-scale. It also overcomes the viral culturing issues because it does not always require in vitro pathogen growth for the peptide expression belonging to viral proteins [24].

Some other advantages of this technique include safety benefits regarding mutations and side effects of attenuated viruses. It is also used as immune agents to activate humoral and cell mediated immune response on a vital domain of viral protein [25] with well-defined synthesized peptides. The central and most important aim of vaccine preparation is to identify epitopes which can generate humoral and cell mediated responses [26, 27]. Due to the presence of both pathogen and host genome sequence, it becomes easier to identify drug targets [20].

In this study, pan genomic approaches, detection of target site and vaccine which is based on peptide epitope is designed to cure and target *clostridium botulinum* will be used. *Clostridium botulinum* has 51 strains that contain 3650 genes collectively. Identification of potential targets for drug development will be done by subtractive genome techniques. Drugs will be analyzed by MIC determination by agar dilution method after in-silico analysis (future perspective).

1.3 Aims and Objectives

The main aim of this research project is to enable us understanding the genomic diversity of *Clostridium botulinum* as well as the identification of novel drug/vaccine targets against this pathogen that may lead to the discovery of novel therapeutics

for the active treatment of botulism in human. To achieve this aim, study is designed with following objectives:

- To explore the pan-genome and essential genome of *Clostridium botulinum.*
- To analyze potential natural compounds with their activity as drug against *Clostridium botulinum.*
- To identify potential virulent factors in *Clostridium botulinum* and evaluation on the potential as drug and vaccine targets.

Chapter 2

Literature Review

This chapter would explain distinguishable features of the research that critically analyze the concept with respect to our research.

2.1 *Clostridium botulinum*

In both marine and surface sediments, *Clostridium botulinum* exists. The *C. botulinum*, which is identified by A-G characters, produces 7 different immunological toxins. Other elated clostridium species such as *Clostridium baratii* and *Clostridium butyricum* are also developed with botulinum toxins and all these toxins are causing highly infective syndromes [28]. Form A, B, E and seldom F of the toxin triggers human infections. All the toxins produced are big, single, structured polypeptides. Toxins damage the presynaptic motor neuron which blocks the transmission of acetylcholine through the neuromuscular intersection of neurons muscles, which results in flaccid paralysis and induces neuromuscular blockades [29, 30].

These contaminants also affect the adrenergic system, but there are no major effects. The widely quoted average lethal dosage of distilled crystal toxin A is 0.09-0.15 mg for man intravenously, 0.80-0.90 mg for man pure botulinum toxin A for oral induction is 70 mg [31]. Nevertheless, some other findings indicate

significantly lower data [32] and these estimates are focused on human cases [33] since the mouse bioassay is the standard tool used to diagnose and quantify toxins. The biological activity of the intraperitoneal lethal dose of the mouse is normally measured by the toxin (MIPLD50).

A spore is produced by the stressed *C. botulinum* and can withstand the normal steps for cooking and food production. Anaerobic medium, poor salt, non-acidic pH and sugar can be hardly achieved in food, and this describes the limited number of foodborne botulisms, which are several other factors for spore germination. A recent, modern industrial canning method called retort canning has been produced for the killing of *C. botulinum* spores [34].

In comparison to spores, botulinum toxins are vulnerable to temperatures as heating all toxins for 5 minutes at 85 C [34] will inactivate them. The germination and vegetation of C. botulinum are not favoured in the typical human stomach. As spores of *C. botulinum* are ingested and expelled normally by people, they do not germinate, produce toxins, or harm the person they pass through. There are a limited number of child botulisms and a small number of adults may experience botulism. Table 5.1 is representing all the present *C. botulinum* strains along with their major details.

2.2 Epidemiology of Botulism Syndromes

Botulism is an uncommon but deadly condition produced by a toxin that affects the body's nerves, causing breathing difficulties, muscle paralysis, and even death. *C. botulinum* as well as *Clostridium butyricum* and *Clostridium baratii* bacteria, produce this toxin. The toxin can be produced by these bacteria in food, wounds, and newborn intestines. Botulinum toxin-producing bacteria can be found in many places naturally, but they rarely cause illness. Spores, which act as protective coverings, are produced by these microorganisms. Spores aid bacteria in surviving in the environment, especially in harsh environments. Even when consumed, the spores normally do not make people sick. However, under specific circumstances.

These spores can proliferate and produce one of the most dangerous toxins ever discovered. Some of the most common signs and symptoms of botulinum includes drooping eyelids, double vision, difficulty in swallowing, slurred speech, muscle weakness, and difficulty in moving eyes etc. Some of the most important botulism types are explained below

2.2.1 Foodborne Botulism

By the consumption of the botulinum toxin contaminated food, foodborne botulism can be caused. *C. botulinum* can only grow and elaborates when the food is present in specific conditions that are important for the *C. botulinum* growth at 4.5 pH, anaerobic milieu, sugar and salt level is kept low and optimum temperature is from 4 C to 121 C [34].

In the continental United States, the major source of intoxication is home-canned foods [35]. The traditional Alaskan indigenous dishes fermented and served without often face a substantial-high risk of food consumption [36]. If the patient has a 3–5-day history of feeding, more proof of the diagnosis of botulism in food can be given that the ingestion of domestically ground food raises the risk of foodborne botulism.

The average number of botulism foodborne cases was 23 per annum in the United Cases from 1990 to 2000, ranging from 17 to 34 cases. Most cases are intermittent, as the outbreaks are usually minor, involving 2 to 3 people and triggered by commercial food or restaurant food [37]. These outbreaks are sporadic [38]. If the patient has a 3–5-day history of feeding, more proof of the diagnosis of botulism in food can be given that the ingestion of domestically ground food raises the risk of foodborne botulism.

Near contact with mutual food may still be observed; this information is vital for the progression of the patent as if the patient were faced with respiratory collapse rather than mechanical ventilation despite the use of supporting and specialized therapy that may also jeopardize the patient's capacity for speech.

2.2.2 Botulism of Wounds

It develops when infected by spores of *C. botulinum* that can be found in the environment, followed by the germination of these spores that can occur in the anaerobic state of the wound. Until the early 1990s, this condition was very rare as after this time, a continuing and dramatic increase of this incidence was experienced by the United States especially among the injection drug users [38].

All these persons that were associated with this injection drug were taking the "black tar heroin" which is a special type of prepared heroin and people take in the "skin-popping" as they inject this drug into the tissues that are opposed to veins [39]. Typical patients of wound botulism are adults that are in their forties or fifties and have a long history of this drug intake and are residents of the western United States. It is very hard to develop an incubation period because most people inject this drug multiple times a day. Clinically this syndrome is indistinguishable from foodborne botulism but it has one difference that is, wound botulism does not develop gastrointestinal symptoms.

2.2.3 Infant Botulism

The ingestion of in situ toxins from colonial *C. botulinum* in the intestines of certain babies 1 year of age is responsible for infant botulism [40]. Infants' botulism this type of botulism is one of the most widespread. There are around 80-100 reports in the US of child botulism [37].

Due to natural bowel florae, colonization takes place and may interact with a not completely developed *C. botulinum*. Some tests have found honey used to be a risk factor for the condition, but just 20% of child botulisms [41] have been reported. In Philadelphia in Pennsylvania, the largest incident of this condition is, however, unexplained [42].

The clinical appearance of child botulism parallels typical signs such as sucking and swallowing, ptosis, wakeful speech, and disappearance of the disorder in adult

types. It can also advance through widespread blackness and respiratory damage [43]. Newly approved antitoxin from human sources is the treatment unique to botulism in infants and reduces the median time from 6 weeks to 3 weeks. With or without antitoxin therapy, the survival rate is around 100 percent if enough intensive treatment is provided [40].

2.2.4 Botulism with Adult Bowel Toxaemia

Adult botulinum toxin ingestion in situ is rarely the source of botulinum toxin invasion in the United States if a few adults are caused by the botulinum toxin that makes Clostridia. The abnormality or antimicrobial use of patients with anatomic or functional abnormality may shield usual Clostridia species from their competitors to normal bowel flora [41].

Due to the continued development of intraluminal toxins, prolonged symptoms and recurrence can be seen in the presence of antitoxin treatment. Proof of long-term excretion and toxin in stools is used for the diagnosis of intermittent botulism in patients.

2.2.5 Inhalational botulism

In 1962, German laboratory staff identified this disease as not naturally occurring and similar symptoms of botulism were present [44]. Inhalational botulism can be caused by the systematic dissemination of botulinum toxin by aerosol [45].

2.2.6 Botulism of Iatrogenic

When botulinum toxin is administered for cosmetic or medical goals, iatrogenic botulism can develop. For systemic disorders, aesthetic care doses are highly suggested. The large doses injected for treatments for muscle motion problems can cause anecdotal cases with systemic symptoms similar to botulism.

2.3 Botulinum Neurotoxin

The gram-positive, spiralizing, anaerobic bacteria belonging to the Clostridium genus have until now been developed by the active botulinum neurotoxins (BoNTs). The causes of flaccid paralysis attributed to *C. botulinum* species are these neurotoxins. *C. Botulinum* is a division between four different classes of heterogeneous bacteria. Most *C. botulinum* species contain only one kind of BoNT and two BNT types are very unusual strains as bivalent strains of 3 types, as is the case with trivalent strains. The groups of C. botulinum types that produced BoNT types are given below in Table 2.1.

TABLE 2.1: Various types of Neurotoxins produced by different strains of C. botulinum

Groups	**Types of strains**
Group I *C. botulinum*	A, H (F/A or H/A) and proteolytic *C. botulinum* B and F strains
Group II *C. botulinum*	E and glycolytic, non-proteolytic *C. botulinum* B and F strains
Group III *C. botulinum*	C, D, CD and DC
Group IV *C.argnentinese*	*C. botulinum* G
Group V *C. baratii*	BoNT/F7
Group VI *C. butyricum*	BoNT/E4 and E5

Groups I to IV of *C. botulinum* strains belong to a genus of bacteria that were differentiated by morphological and genetic differences that correspond to different bacterial levels. The differentiation of different clostridium species that have BnNTs is confirmed by whole genome sequencing and phylogenetic analysis.

In group, I strain the name C. para botulinum was suggested, in group II strains the name *C. botulinum*, in group 3 strains the name C. novyi sensu lato and in group IV it is the name *C. argentinense* [46].The Group III C strains Botulinum contains, especially in the phages not incorporated into chromosomes, neurotoxin

genes located in bontC and bontD [47]. Over the C. Cultivation of botulinum, the phage of bont-haven can quickly be lost and is not poisonous anymore. The III C party.

The non-toxicant derivative strains of bontC and bontD are also commonly discovered as non-toxic derivatives that are derived from biological or environmental samples. The species of non-neurotoxigenic clostridium are associated by both the groups. *Clostridium haemolyticum* and *Clostridium novyi* are similar to group III of *C. botulinum*.

C and D phages include bontC, bontD and plasmids of the neurotoxin genes interchangeable between category III *C. botulinum*, *Clostridium haemolyticum* and *Clostridium Novyi* [48].

2.4 Animal Botulism

2.4.1 Group III Animal and *C. Botulism*

A to G groups will invade any spinal cell because of the absence of unique receptors on the neuronal cell surface [48] of the invertebrates showing resistance to BoNT/A to G. Human beings and all vertebrates are susceptible to advances of botulism.

This sensitivity varies, however, depending on the BoNT class, and it also depends on the presence of botulism receptors on such a large scale on the body of the neuronal cells and the presence of intracellular cleavable target isoforms (SNAP25, VAMP). A novel BoNT that is known as the paraclostridial mosquitocidal protein1 (PMP1) is found recently that is produced by the Paraclostridium bifermentans subsp Malaysia strain that is Anopheles mosquitos [49].

This BoNT is the first kind of neurotoxin targeting species of invertebrates. The new type of strain known as Bengtson's and Seddon's type C of 1992 is described in the United States as strains that were isolated from the chickens with botulism, and in Australia strains are isolated from the botulistic bovine strain [49]. The

Bovine, which had Meyer and Gunnison botulism in 928, has been isolated from a distinct microorganism in South Africa and the BoNT is referred to as D Type [39].

Type C and D were more often present in animal botulism and then BoNT/DC and BoNT/CD mosaics [50] and owing to their biochemical properties these species have been assigned to *C. botulinum* III. As the optimal temperature for this community is 40c, these strains can occur in the warmer regions and the warm seasons in the moderate regions around the globe [49].

2.4.2 Group III Presence of *C. Botulism* in Birds & Cattle

Botulism, which causes significant economic losses worldwide each year, is predominant in birds and livestock. It is more intermittent in the other animals. Animals are seen as although a category C and D botulism, and various animals are different to the BoNT/C and BoNT/D sensitivities. E.g., the birds are BoNT/D resistive and have a BoNT/C sensitivity while the animals are both type C and type D sensitive. The C type is much more sensitive than bird ducks, pigs, fheasans, and turkeys [50]. Notice that certain special BoNT samples can be neutralized, including anti-C and anti- /D serums. The C. botulinum C/D BoNT is used both to neutralize normal anti-BeNT/C and anti-BoNT/D sera, and only for the anti-BoNT/D serum BoNT.

Mostly wild birds like ducks and birds on farms are affected [51]. Because of this, routine out brings in aquatic birds in Europe, Japan, and North America are often recorded in essential botulism [41]. In wild birds, for example, 13 outbreaks of botulism causing the death of more than 50,000 birds have been recorded. It takes place from 978 to 2008 [52]. There were 129 outbreaks of botulism in wild birds and 396 farmed birds in France from 2000 to 2013 [53].

In a study of 17 flocks populistically affected, the mortality rate of each flock was observed at between 10,000 and 20,000 farmed birds at 2.8 to 35 percent [54]. Between 2000 and 2008, the deaths of around 68,000 wild birds occurred [55]. For

example, Botulism in fowl is common of BNT or organ type C or D type examined by molecular biology in blood or organ samples [56].

2.5 Toxin Structure

2.5.1 Complex Toxins

Both normal and in vitro cultures have botulinum toxins coupled with other non-toxin proteins [55]. The complex of toxins of C, E and F and the C. botulinum hemagglutinin-negative strain D involves the movement of the neurotoxin with comparable molecular weight, i.e. 150 kDa, the interaction of other non-toxic protein [35]. Based on the type of toxin that can range from 230 to 350 kDa, these M-complexes have varying molecular weight.

The type A, B and haemagglutinin-positive strains of type D may contain a third protein component that exerts hemagglutinin action, greater than 450 to 500 kDa, like an L complex. The molecular weight of the toxin type A is about 900kDa, which may contain L part dimers. Multiple moieties of neurotoxins include various serological properties. Structural comparisons also exist between types E and F M of the non-toxic elements of type C and D M complex [55].

The protein complexes dissociate at high pH levels and can be pH 8 but can spontaneously reform at lower pH. Hemagglutinin activities may be present in certain non-toxin proteins and their roles are largely uncertain. Nevertheless, the neurotoxin is protected from the aggressive intestinal setting. In the circulatory and lymph system, it improves absorption. Compared to purified neurotoxins, the botulinum toxin complexes are more toxic when delivered by mouth than neurotoxins that are parenterally provided with particular toxicity higher than the toxin complex [5]. These toxins have a comparable molecular Weight between 140-170kDa and are mainly synthesized as single polypeptide chains, and all the neurotoxins except for G toxin have been purified to harmony [55]. They have two subunits which are asymmetric and consists of 85-105 kDa(heavy chain) and

50-59 kDa(light chain) respectively which is derived from the proteolytic active form of protoxin and joined together by at least one disulfide bridge [57] in the most active form of the toxins. The other toxin forms, such as A, B and E, the amino-terminal of the polypeptide chain have been found from the most extensive research.

C. botulinum neurotoxin fragments of type A have recently been colonized with E. coli. The full fusion amino acid sequence is now known, [55] and the polypeptide of 149.580Da is present in this predicted sequence. The study of this sequence will expand awareness of potential functional features of molecular aspects of the molecular structure.

The proteolytic cleavage site which comprises the active toxins derived from the protoxin includes functional aspects known. The regions that have a high hydrophobicity potential and are linked to the observed in-vitro canal shaped behaviour are also identified [55].

This known sequence also helps to forecast the secondary molecule composition. The tetanus toxin and botulinum neurotoxin have some correlations in the molecular structure and basic modes of action [55]. Tetanus toxins look structurally identical to botulinum neurotoxin in molecular size and composition. By gene cloning, the full tetanus toxin amino acid sequence has been obtained [58] and a comparison of botulinum type A toxin [55] to strongly preserved regions in the N-terminus area of heavy chains was expected to have a 33-percent homology. In future, gene cloning will certainly contribute to clarifying the genomes of other botulinum neurotoxins and will help to identify large structural similarities and discrepancies between various serotypes of the toxins.

2.5.2 Mode of Action

The toxicities of the neurotoxins vary from 2 to 10 to 2 — 10 mouse LDs protein doses mg-1 and are the most active agents to inhibit the acetylcholine neurotransmitter by functioning presynaptically at the neuromuscular crossing. In this

inhibition, there are three main stages [55] that include in Fig 2.1.

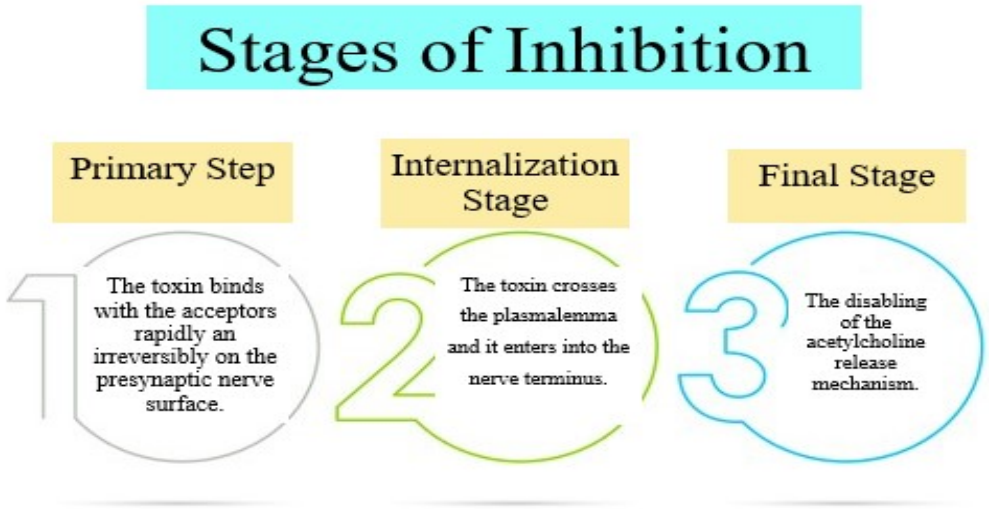


FIGURE 2.1: Stages of inhibition of *C. botulinum* in host organism

Neurotoxins have been shown to attach selectively and saturably to the presynaptic neuronal membranes of both peripheral and central nerves [59]. But the possible characteristics of the receptor molecule on the presynaptic surface that bind with the neurotoxins are not yet proved and also it is a fact that all the toxins do not bind with the same acceptors.

Toxin types A and B do not compete for the same acceptor site in rat brain synaptosomes as a nerve model, however toxin types A and E compete for the same acceptor because they share the same acceptor [60], and toxin types C and D also contend for the same acceptor site [61].

Neurotoxins were selectively and saturable binding to both peripheral and core membranes of the presynaptic nerves [59]. However, it has not yet been shown that all toxins bind to the same acceptors and that the potential properties of the receptor molecule on the presynaptic surface bind with the neurotoxins are not proven to be. Studies on the brains of the Rats as a nerve model reveal that Type A and Type B toxin do not interact at a single site, while Type A and E toxin compete with the same site as an acceptor [60], and Type C and Type D toxins also show up at the same site [61]. The associated poison accepters A, E or B are

also unidentified. Two small-scale site-binding neurotoxin receivers (KD) appear to have a high-binding affinity (kd¡lnM). The neurotoxin can be required to bind to such acceptors from an active binding area at the carbonyl terminal, which is half the heavy subunit. Silicium acid residues are potentially trisialogangliosides and proteins in neurotoxins [55].

2.5.3 Internalization

After the binding, the acceptor that is present on the nerve surface of the neurotoxin is internalized through an energy-dependent process. By using the mouse hemidiaphragm of the mouse that was prepared through the incubation with the radio-labelled neurotoxins, direct evidence for internalization was obtained [55]. After 90 rain at 22 C, the internalization with the A and B toxins was maximal.

The receptor-medium endocytosis is analogous to this process, except the molecules are enclosed in clathrin-coated endosomes, where the interval state can turn acidic as they go through the cytosol. The heavy-chain mobility of the neurotoxin induces channel formation at low pH in lipids [55]. But the whole toxin or its fragments are unknown and they are delivered to the cytosol through the process of internalization process. According to recent studies, it is suggested that the internalization of both the heavy and light neurotoxin chains is required for neurotransmitter release inhibition.

2.5.4 Neurotransmitter Release

The neurotoxin's activity is inhibited when one or more stages of the neurotransmitter's release mechanism have been integrated. The pathway is not impaired by neurotoxins but acts on the activities following the influx of the calcium ion. Some possible inhibition mechanisms are proposed are given in Fig 2.2 [62]. The present data is insufficient to prove any above-given mechanism but it is believed that the botulinum toxins can block the release of the transmitter by common mechanisms [55] [63].

It is also believed that the toxins exert their toxicity by some enzymic activity but the mechanism is still unknown [55] [60].

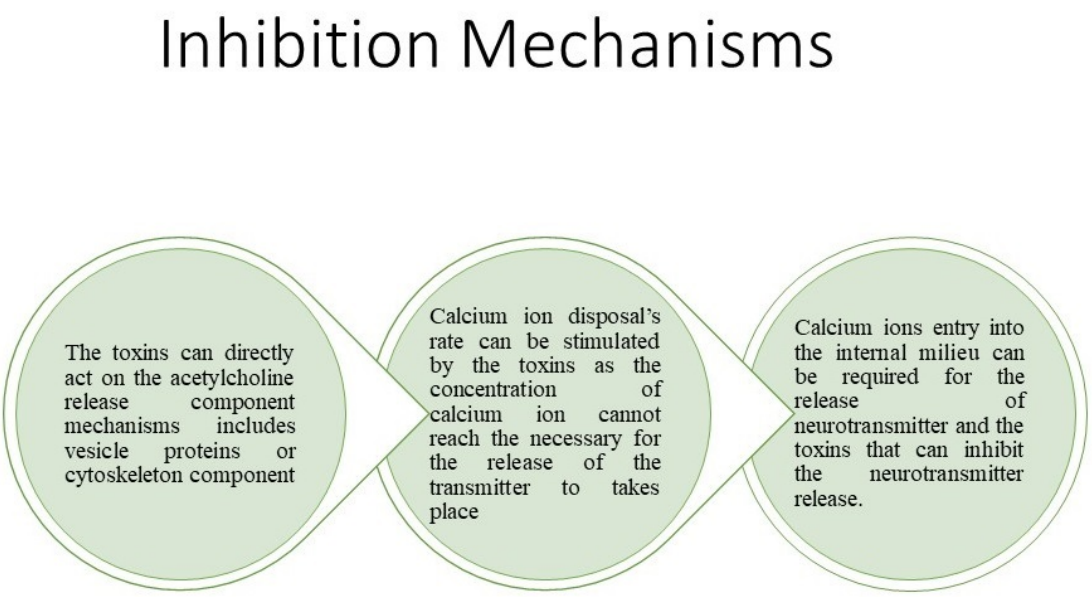


FIGURE 2.2: Possible Inhibition mechanisms of *C. botulinum*.

2.6 Pan-Genome Analysis of *Clostridium botulinum*

It is the whole set of genes from all the obtained strains present in a group containing the same ancestor in molecular biology and genetics. In a broader sense, it's the sum of a clade's genomes. The pan-genome is divided into three sections: a "core pangenome" that contains genes found in all individuals, a "shell pangenome" that comprises genes found in two or more strains, and a "cloud pangenome" that only contains genes found in one strain. Some writers refer to the cloud genome as an "accessory genome," which contains 'dispensable' genes found in only a fraction of strains as well as strain-specific genes. The advances in the technologies of next-generation sequencing have revolutionized the understanding regarding cellular localization, functional diversity at the metagenomic level [64] and microbial genetic repertoire [65]. In addition to this, whole-genome sequencing of bacterial pathogens helps in prioritizing the insert of researchers towards the pathogenicity by accurately measuring the genetic variations among the pathogenic groups [37].

The cost-effective and time-consuming identification of variable sites referred to as SNPs, is used to infer genetic variants among numerous genomes at the bench-top level. The whole-genome multi-locus sequence typing (MLST) technique [66] can be used to accomplish this purpose. To overcome the possible limitations of these reference-based methodologies, a comparative genomic method is applied. This method is based on sequence similarity search analysis [2], and it is causing a shift in worldwide attention in omics methods.

These can be done by the whole-genome multi-locus sequence typing (MLST) approach [66]. A comparative genomic approach is used to reduce and sort out the limitations that are relevant to these approaches which are based on references. This approach depends upon the analysis of search of similar sequences [2] and it shifts the global interest towards the omics strategies. The foundation of omics is laid by the availability of sequenced data at the public repositories and its free accesses and it also helps in the formation of the consecutive system biology principles [67].

The comparative microbial genomics strategy that is based on the sequence similarity helps in identifying the essential genetic content that is shared by all the pathogenic isolates with the help of statistical analysis. It also helps in finding the genes that encode virulence and novel functions as a variable genome [68].

Both core and variable genome content of an organism is signified by the pan-genome [69, 70] and the whole genetic repertoire of the isolates are represented by the supra-genome. A phylogenomic study using the pan-genome assists in determining the genomic contents of a group of organisms, such as the core, pan, and variable genomes, as well as their lifestyle, which can be allopatric or sympatric [69].

C. botulinum is a gram-positive, anaerobic bacteria and it produces widely existable heat resistant spores [67]. As it is an anaerobic bacteria due to which it produces spores that germinate, grow and secrete toxins in the absence of oxygen [37]. These toxins are the most potent a third most infectious agents that offer globally a greater risk to animal and human health.

According to its global distribution, it is found out that in selected locations in India and other developing countries it acts as an endemic [71]. From a range of host-pathogenic interactions, its pathogenicity, toxicity, colonization, transmission from one organism to another and invasiveness [72]. The mortality rate kept on increasing despite the research and improvements in the therapies [73].To find and priorities therapeutic targets against *C. botulinum*, a computational framework and a silico system-level framework have been established. In humans, *C. botulinum* is the cause of flaccid paralysis. It can be fatal in 5-10% of instances, and because of the emergence of drug-resistant infections and its scarcity, the disease is extremely difficult to control, as there is only one antitoxin that can target neurotoxic at the extracellular level. To this end, an exhaustive system size study has been conducted of genomic sequence homology, phylogenetic connections between *C. botulinum* and other infectious bacteria with a host or human intestinal flora [74].

2.7 B-Cell Epitope Mapping

It is a potential method for finding microbes' key antigenic determinants, particularly those with discontinuous conformational properties. Epitope-based vaccinations have several advantages over traditional vaccinations, including being more precise, avoiding unwanted immune reactions, generating long-lasting protection, and being less expensive. The essential to antigen-antibody interactions is the humoral immune reaction to invading pathogen. A particular antigen (Ag) is recognized as the antigen determinant of B-cell epitopes by the particular antibody (Ab) in the distinct region. Surface available clusters of amino acids recognized as B cell epitopes by secreted antibodies. These epitopes of the B-cells may produce cell or humoral immune response [75]. Many surface antigens may be used in the epitopes after antibodies have been recognized. This method is not yet known about the mechanism [76]. By lacking the antigen reconfiguration in the Ag-Ab complex, the idea of classifying antigen in epitopes and non-epitopes is not capable of reliably reflecting the bioactivity [77]. Precise identification of the B-cell

epitopes [77] is another immune diagnostic technique [63] which is the basis of the advancement of antibody therapeutics and vaccines which are based on peptides.

It can be categorized as a nonlinear and continuous conformational structure, depending on the spatial structure. Amino acids are in closest interaction with the discontinuous epitopes because of the three-dimensional conformation [63]. A minimal amino acid sequence is expected of the indigenous proteins for the proper folding of discontinuous epitopes. The range of the amino acid series is from 20 to 400. The bulk of the linear antigenic determinants found are thought to be the pieces of the conformational B-cell epitopes [78]. Analysis has shown that more than 70 percent of discontinuous epitopes consist of 1 to 5 linear parts. The length of these segments can be of amino acids varying from 1 to 6 [78].

The methods for determining the epitopes can be developed experimentally, hence can be separated as functional and structural studies approximately. The precise position of the laborious can be located precisely by the crystallography of X-rays which is a technique that takes time and is technically hard. This procedure therefore should not apply to all antigens [78]. The most popular approaches used for functional mapping of beta cells include proteolytic fragments screening extracted from antigen for the binding of antibody and measuring the reactivity of mutants with Ag-Ab. Site-driven or spontaneously mutating are these mutants [78].

Any of these techniques are inexpensive, fast and versatile in comparison to other techniques, and are used in the study of the epitope mapping [79]. An antigenic surface is homogeneously antigenic, according to which Rubinstein and his colleagues have established a null hypothesis. The epitopes were also described in the form of broad statistical analyses of Ag-Abco crystals deriving from protein databases to describe the physicochemical, structural and geometrical aspects. With this data, it can be inferred that the epitopes can be distinguished from the other surface of the antigen [78].

Another research was carried out by the Kringelum and colleagues who identify the smooth, extended, oval-shaped bundles which have an unknown secondary

structure as the B-cell epitope [80]. The distinguishable characteristics of epitopes and non-epitopes are recognized using systematic laboratory studies and silicon analysis. The bulk of the epitopes are between 15 and 20 residues and are arranged into loops between 600 and 1000 to 1000 A2. The epitope's surface usability is the most common function. The epitopes series is complemented by Y, W, loaded polar amino acids. Relevant amino acid pairs are also available. The relationship between Ag-Ab and complex CDR loops involves epitope compression [78].

In recent years, the distinctions between residues between peptides and other residues are not taken into account [76]. Advances in epitope mapping technologies and bioinformatics are very important for immune technology creation and include uses of the numerical approach to disclose anticorps, B-cells, T-cells, and other allergen structures. The results will also be discussed. [81]. In the early prediction methods, the linear epitopes were identified by the propensity stage. Multiple techniques, like machine study methods like the Hidden Markov Model [82], supporting vector machines and the recurrent neural network [83], have been introduced to improve prediction efficiency. Despite these advances, there is only a small range of approaches used to forecast discontinuous epitopes. These approaches need to combine knowledge such as statistics for amino acids, sensitivity to surface areas and spatial data [84].

B-cell epitope discovery is very significant in advancement for testing of diagnostic therapeutics and vaccines [55]. Epitope imaging has been used in drug development[85]. Examinations are underway. Despite getting results in the mapping of B-cell epitope, this is also essential for observing antibodies against peptides because they cannot bind to native proteins because of the unstructured nature of peptides [86].

2.8 Virulome

Virulome is "The set of genes that contribute to the virulence of a bacterium". Microbial pathogenesis is determined by the virulome secretion machines, secreted

virulence effector molecules and common regulatory mechanisms [76]. Gene products involve virulence factors of the bacterial toxins and the cell surface proteins which help the pathogens to develop themselves into the host [86]. It is very imperative to control the hazards of foodborne botulism by understanding the molecular hypothesis behind the pathophysiological mechanisms of the type A strain. This strain can produce neurotoxins and other virulence factors that can determine the potential and degree of molecular virulence of this type of neurotoxin in the human gut. As foodborne botulism is not caused by the infection, due to which the virulence study and metabolic cross-talk of it on a genome-scale help in the recent systems biology research. The virulence factors that are produced from the pathogens and the pathogen-host-interactions that are involved in the disease conditions are comprehended by the pathogenic studies [86].

The botulinum toxin is used by the *C. botulinum* to kill the host for subsequent saprophytic utilization. *C. botulinum* does not spend long periods during their association with a living host. Type A strain contains 2adherence proteins that are fbp and groEL while three toxin coding genes in addition to the BoNT/A that are cloSI, colA, and hlyA [75].

The virulence factor-coding genes are located on the chromosomes of the type A strains but the type A3 strain contains its virulence factor-coding gene that is BoNT/A in the plasmid. The operome preliminary analysis found that the Hall A strain consists of 6 new virulence factors that include Gp11 baseplate wedge subunit and tail pin, arginosuccinate synthase, internalin A, adipocyte enhancer-binding protein 1, C-junamino-terminal kinase-interacting protein 4 and von Willebrand antigen II. This indicates the many obscure mechanisms for its pathogenicity mechanisms. The components of the bacteriophage are the Gp11 baseplate wedge subunit and tail pin, it suggests that all the virulence factors may not be e threat to humans and some of these factors can be only virulent towards the other bacterium. There are no present studies that can clear the reasons behind the production of these virulence factors in the first place and the mechanisms that show the entry and survival of these factors in the host. It is also obscure that how these factors find their target and perform their toxic activities.

Chapter 3

Materials And Methods

3.1 Drug Targeting Techniques

Potential drug targets were found out by multiple-step methodology. This methodology includes genome selection, identification of core genome, non-homologues protein identification, target identification, catalytic pocket detection and molecular docking.

3.1.1 Genome Selection

The whole genome of *C. botulinum* was searched and obtained from the National Centre of Biotechnology Information (https://www.ncbi.nlm.nih.gov) [87]. 51 strains of *C. botulinum* were used in this study. Table 5.1 is showing all the strains of *C. botulinum*.

3.1.2 Identification of Core Genome

For the identification of the core genome, EDGAR 2.3 (https://edgar. computational. bio.uni-giessen.de/cgi-bin/edgar.cgi) was used. From all the 51 strains core genes were retrieved by using Clostridium botulinum A str. ATCC 3502 as

a reference strain based on its release date. Genes that were common in all the strains were selected for further analysis. In Edgar 2.3, one strain was selected as a reference strain that was Clostridium botulinum A str. ATCC 3502 that was compared with all other 50 strains [88].

3.1.3 Non-homologous Protein Identification

After the retrieval of the core genome through the EDGAR, non-homologous genes/proteins were identified through the BlastP (https://blast.ncbi.nlm.nih.gov /Blast.cgi?PAGE=Proteins). These non-homologous proteins were inserted into the DEG Database (http://tubic.tju.edu.cn/deg/) to find out essential proteins [88]. From the set of proteins, by applying thresholds i.e. identity ¿35 and e-value=0.001, 9 proteins were selected from which 7 cytoplasmic proteins were chosen based on cellular localization [89] for drug targeting study and 2 were selected for the epitope-based study.

3.1.4 Target Identification

For the determination of potential therapeutics, multiple factors such as molecular weight, pathway analysis, virulence etc. were used [90]. Multiple protocols were applied on the essential genes such as molecular weight identified by ProtParam (https://web.expasy.org/protparam/) [91].

Kyoto Encyclopedia of Genes and Genomes (KEGG https://www.genome.jp/ kegg/?sess=ebfe2ad23e021e38540f7 98 c803dd061) was used for the pathway analysis [92].

Virulence was found out by Pathogenicity Island Database (PAIDB http://www. paidb.re.kr/about-paidb.php), molecular and biological process was found by Uniport (https://www.uniprot.org/) and CELLO: Subcellular localization predictive system (http://cello.life.nctu.edu. tw/) was used for finding the cytoplasmic and membrane proteins [93].

3.1.5 Catalytic Pocket Detection

DoGSiteScorer (https://proteins.plus/) was used to find out the catalytic pocket of the essential proteins with the specific drug score [94]. DoGSiteScorer is an automated tool that is used for pocket detection and to calculate the drugability of protein cavities.

3D model or PDB id of protein can be given as input for the analysis. The drugability score can range from 0 to 1. A drugability score of more than 0.60 is considered, but if it is above 0.80 then it is more favored [94].

3.1.6 Molecular Docking

MOE [95] was used for the molecular docking with the list of the ligand of anti-bacterial compounds taken from the literature review [96, 97]. Both the minimized and prepared ligands and proteins were subjected to docking by keeping all the default parameters. Best interacted ligands were selected based on binding/S energy and number of interactions. One best-interacted ligand was selected for PyMole analysis.

3.1.7 Docking Validation

For the validation of docking results, RMSD was calculated by MOE and further analyzed by PyMole.

3.2 Epitope-Based Vaccine Identification

Epitope-based vaccine target identification involves a list of steps that includes data retrieval and structural analysis, prediction of B-cell and T-cell epitope, Imperative features profiling of selected T cell epitopes, Epitope conservation analysis, Multi-epitope vaccine design and construction, Physiochemical analysis of

Multi-epitope vaccines, Multi-epitope vaccines 3D Structure prediction and Molecular Docking.

3.2.1 Data Retrieval and Structural Analysis

The chemical and physical properties of the proteins known as a pre-protein translocase, SecG subunit (Mur ligase family protein), and UDP-N-acetylmuramyl tripeptide synthetase-like protein are analyzed using Protparam(https://web. expasy. org/ protparam/). The half-life of the proteins, GRAVY (Grand Average of Hydrophathicity), molecular weight, amino acid atomic composition, and stability index are all physical properties [98]. PSIPRED [99] was utilized to analyze the secondary structure of both proteins (http://bioinf.cs.ucl.ac.uk/psipred/). The transmembrane topology analysis was performed using the online tool TMHMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/). The presence of disulphide linkages in both proteins was analyzed using DIANNA V1.1 (http://clavius.bc.edu/ clotelab/DiANNA/) [100], an online tool for making predictions based on a trained neural system. Vaxijen v2.0 (http://www.ddg-pharmfac.net/ vaxijen/VaxiJen/VaxiJen.html) [101] was used to check antigenicity, whereas AllerTOP v2.0 (https://www.ddg-pharmfac.net/AllerTOP/AllerTOP.html) was used to check allergenicity) [102].

3.2.2 B-cell Epitope Prediction

For the B-cell epitope forecast, the freely available online tool BCPRED (http:// ailab -projects1.ist.psu.edu:8080/bcpred/predict.html) was used [103]. Criteria for B-cell epitope prediction was set as 14 residue lengthy epitopes and 75% specificity[109]. For checking the antigenicity of the resulted epitopes Vaxijen v2.0 [101] was used and only antigenic epitopes used for further analysis. There are multiple factors on which B-cell epitope recognition depends. Antigenicity, surface accessibility, flexibility, linear epitope prediction, and hydrophobicity are among these factors [104].

All of these factors, including the Kolaskar and Tongaonkar antigenicity scales, Parker hydrophilicity prediction algorithms, Karplus and Schulz flexibility prediction tool, Emini surface accessibility prediction tool [105], and Chou and Fasman beta-turn prediction algorithm [106], were performed using IEDB analysis resources (http://tools.iedb.org/bcell/). Because discontinuous epitopes have a greater number of explicit properties than linear epitopes, the DiscoTope server (http://www.cbs.dtu.dk/services/DiscoTope/) [107] was used to predict them. The parameters for this study were set at 0.5, which is 90 percent specificity and 23 percent sensitivity. PepSurf (http://pepitope.tau.ac.il/) was used to analyze 3D protein structures, which aids in indicating epitope sites on the structure.

3.2.3 T-cell Epitope Prediction

CTLs (cytotoxic T-lymphocytes) play a significant part in vaccine development since they save money and time when compared to wet lab research. The MHC class-I and MHC class-II CTL EPITOPES were identified using two online programs, Propred-1 (http://crdd.osdd.net/raghava/propred1/) [108] and Propred (http://crdd.osdd.net/raghava/propred/) [109]. These tools use several human-leukocyte-antigens (HLAs) alleles due to which their result is quite substantial. The threshold for Propred-1 was kept at 5% and for propred, it was 4 %. Proteasome and Immuno-proteasome filters were kept on. Only antigenic epitopes will be used for further analysis.

3.2.4 Imperative Features Profiling of Selected T Cell Epitopes

Important features of chosen epitopes for both MHC class-I and MHC class-II of both proteins were checked that includes enzyme digestion, toxicity, mutation, hydrophobicity, hydrophilicity, charge, PI and molecular weight. Peptide cutter (https://web.expasy.org/peptide-cutter/) was used to find out non-digesting enzymes. For the allergenicity prediction of epitopes an online tool AllergenFP

1.0 [110] (https://ddg-pharmfac.net/AllergenFP/) was used. For toxicity, hydrophobicity, charge, PI, hydrophilicity analyses, ToxinPred (http:// crdd.osdd.net/raghava/ toxinpred/) was used. Molecular weight was found by Protparam. Only non-toxic epitopes were used for further analysis [111].

3.2.5 Epitope Conservation Analysis

Distinctive sequences of multiple proteins of the same species were chosen for both proteins and multiple sequence alignment was performed. Conservation of all the B-cell, MHC class-I and MHC class-II epitopes were checked against all chosen proteins through IEDB conservation-analysis-tool (http://tools.iedb.org/conservancy) [111].

3.2.6 Multi-epitope Vaccine Design and Construction

Multi-epitope vaccines were designed by arranging the epitopes according to their arrangements in the reference sequence. MHC Class-I epitopes were joined through GGGS flexible linker [112] while MHC Class-II epitopes were joined through flexible GPGPG linker [113].

The second multi-epitope vaccine was joined by joining the B-defensin, an adjuvant, was added at the N terminus of multi-epitope with EAAAK linker [114].

3.2.7 Physiochemical Analysis of Multi-epitope Vaccines

Physiochemical properties such as antigenicity through Vaxijen v2.0, allergenicity through an online tool AllergenFP 1.0, protein solubility was checked through an online tool known as Protein-Sol (https://proteinsol.manchester.ac.uk/cgibin/solubility/ sequenceprediction.php).

Other physicochemical properties such as GRAVY, instability index and aliphatic index molecular weight, PI etc. were found through protparam tool.

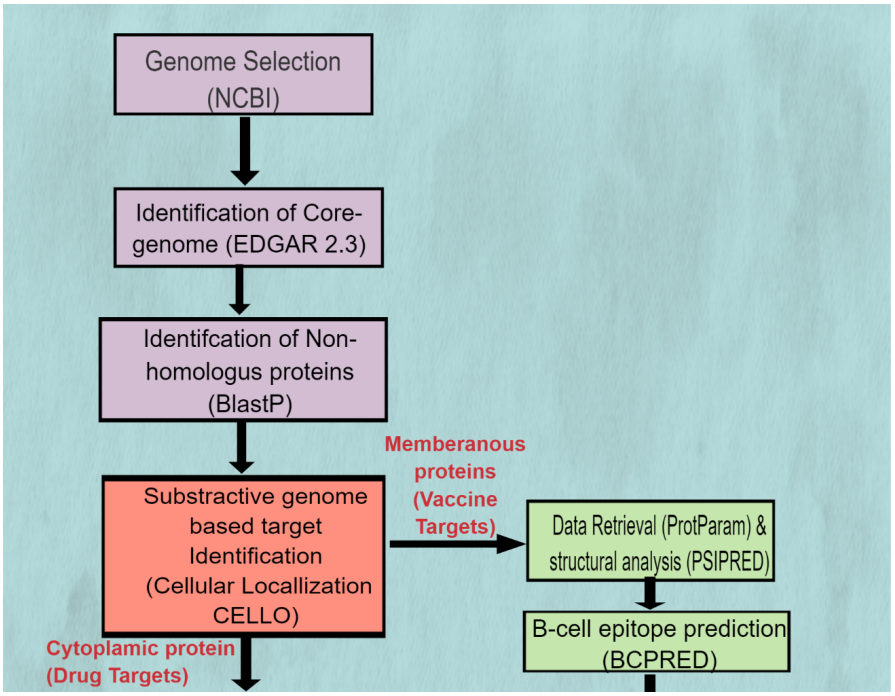
3.2.8 Multi-epitope Vaccines 3D Structure Prediction

3D structures of both vaccines were predicted through Swiss-model (https:// swiss-model .expasy.org/). It gave multiple models based on the sequence but only one top model for each vaccine was selected based on sequence identity in PDB format.

3.2.9 Molecular Docking

The predicted 3-D structures of both vaccines were docked against multiple HLA types that include HLA-A1 (3BO8), HLA-A2 (6AT5.1.A), HLA-A*0201 (1BOR), HLA-A*0205 (5IPF). 3-D structures of all the proteins were downloaded from the PDB RCSB (https://www.rcsb.org/). Docking was done through the Cluspro 2.0 (https://cluspro.bu.edu/publications.php) which is an online tool. PDB* file for Vaccine is given as ligand for input while protein was used as a receptor. The selected model was submitted to PDBsum (http://www.biochem.ucl.ac.uk/bsm / pdbsum) for interaction analysis.

3.3 Methodology Overview



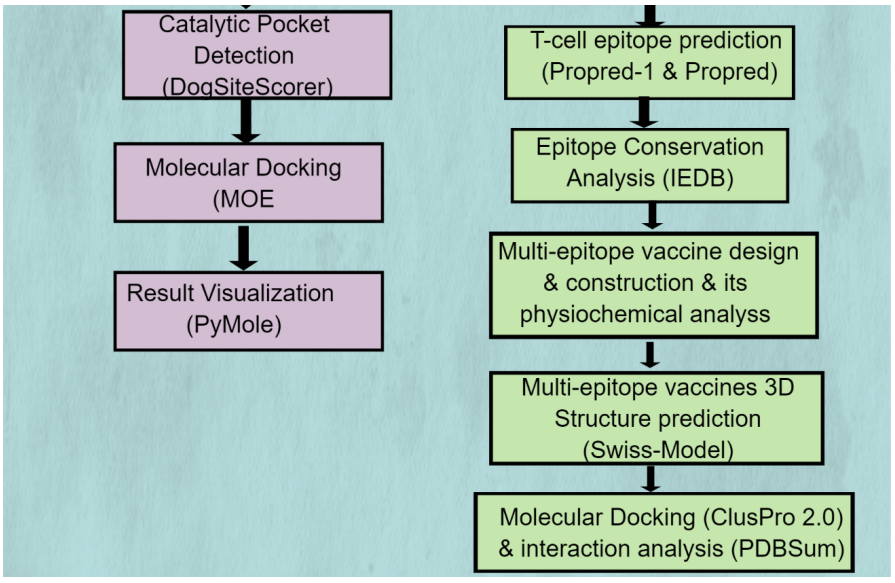


FIGURE 3.1: Methodology used for the drug-targeting and epitope-based study

Chapter 4

Results and Discussion

This chapter will explain the results that were obtained by following our methodological steps.

4.1 Drug Targeting Analysis

From the 51 strains of C. botulinum, our methodology allowed us to find out 7 target proteins that can be used as drug targets for all these strains. All these 3D structures of proteins were validated by RAMPAGE that stands for RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression and ERRAT for validation that is shown in Table 4.1. ERRAT is a non-bonded atomic interaction's "overall quality factor," with higher scores signifying higher quality. The score greater than equal to 80 is considered good for Rampage while greater than equal to 37% is considered good for ERRAT.

After finding all the drug target subtractive genomic-based analysis was performed to find out the molecular weight, pathways in which they are involved, virulence, molecular and biological functions and their cellular localization. From the total 8756 core proteins, there were 422 non-host homologous proteins from which we selected 7 essential proteins by applying 2 thresholds that are identity greater than 35 and e-value = 0.001.

TABLE 4.1: Validation scores of the selected proteins against Rampage and ERRAT

Protein name	**Rampage***	**ERRAT****
Chromosomal replication initiator protein DnaA	91.1%	94.2598
ThiI domain-containing protein	90.0%	89.1304
ATP synthase subunit c	94.8%	85.7143
Putative guanosine 3',5'-bis-pyrophosphate (PpGpp) synthesis /degradation protein	96.4%	94.0541
Putative metal-dependent phosphohydrolase	93.0%	89.5349
UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D-alanine ligase	94.1%	85.3273
Chemotaxis protein	85.0%	81.5385

*Rampage: percentage above greater than or equal to 80 is considered to be a high-quality model for drug targeting studies.

**ERRAT: Range greater than or equal to 37% is acceptable for a high-quality model.

Table 5.4 is representing all the important features of all the essential proteins. All these characteristics of the proteins help to find out the importance of these non-homologous essential proteins for the organism. These are the proteins that are present in all the strains of the *C. botulinum* as they were taken from the core genome of the organism that is shared by all the 51 strains.

4.2 Molecular Docking

Catalytic pocket detection was performed by the DogSiteScorer. Pockets with the highest drug score were selected for each protein. 105 anti-bacterial compounds

were taken from the literature review and prepared by ChemBioDraw Ultra 11.0.

After that these compounds were prepared by MOE by adding hydrogen and minimizing energy for using these compounds as ligands. PDB structures of proteins were taken from SwissProt Expassy also prepared by MOE by protonating 3D and by doing energy minimization. For the validation of the MOE Dock program, RMSD value was calculated.

For this purpose, the crystallized ligand was removed from the protein ATP synthase subunit C (PDB ID: 2HLD) The found RMSD value was 0.9856 A. Figure 4.1 is showing the validity of our docking method. It proves that MOE is reliable for docking studies.

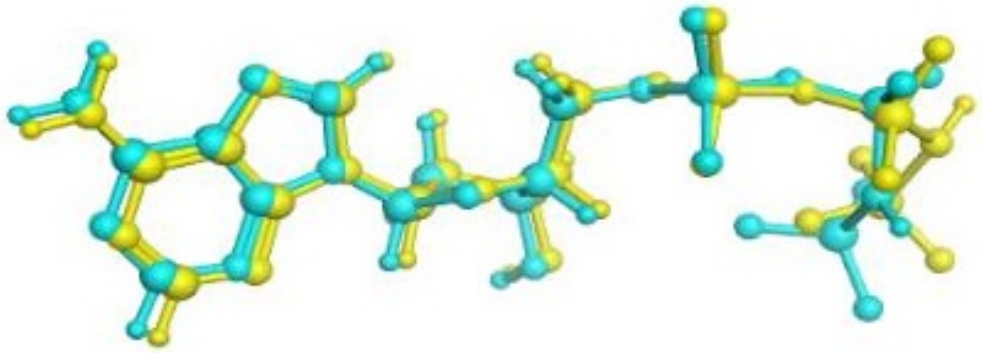


FIGURE 4.1: Docking of multiple ligands

After that, each protein is docked against all the prepared ligands and the top 10 compounds with the highest binding affinity were chosen for all the proteins.

4.2.1 ATP Synthase Subunit C

The docking results of ATP synthase subunit C are explained below. It is a cytoplasmic protein that is synthesized by ATP E. Its molecular function is that it is involved in the lipid-binding and proton-transporting ATP synthase activity. Biologically it also acts as ATP synthesis coupled proton transported and has 77.6 kDa molecular weight.

TABLE 4.2: Compounds name, binding affinity and Residue interaction of top 10 compounds with ATP synthase subunit C

Compounds	**S-Score/ binding affinity***	**Residues interaction****	**No of interactions*****
2,3-dibromo-5,8-dihydroxynaphthalene-1,4-dione	-8.9291	THR 64	1
1,8-dihydroxy-3-methoxy-6-methylanthracene-9,10-dione	-8.8235	LYS35	1
5,8-dihydroxy-2,3-dihydronaphthalene-1,4-dione	-8.3052	LYS35	1
(E)-4-(4-bromo-2,5-dimethoxyphenyl)-4-oxobut-2-enoic acid	-7.8215	2/ LYS35	2
2-bromo-5,8-dihydroxynaphthalene-1,4-dione	-7.7162	THR64	1
2-bromo-5-hydroxynaphthalene-1,4-dione	-7.7102	THR64	1
Juglone	-7.4304	LYS35	1
5-hydroxynaphthalene-1,4-dione	-7.3346	LYS35	1
2-(3-methylbut-2-enyl)-benzene-1,4-diol	-7.0672	LYS35	1
Plumbagin	-6.9453	2/LYS35	1

*The intensity of the binding relationship between a protein and compound (ligand)

** Residues that are involved in the interaction

*** No of interactions formed between active site and ligand

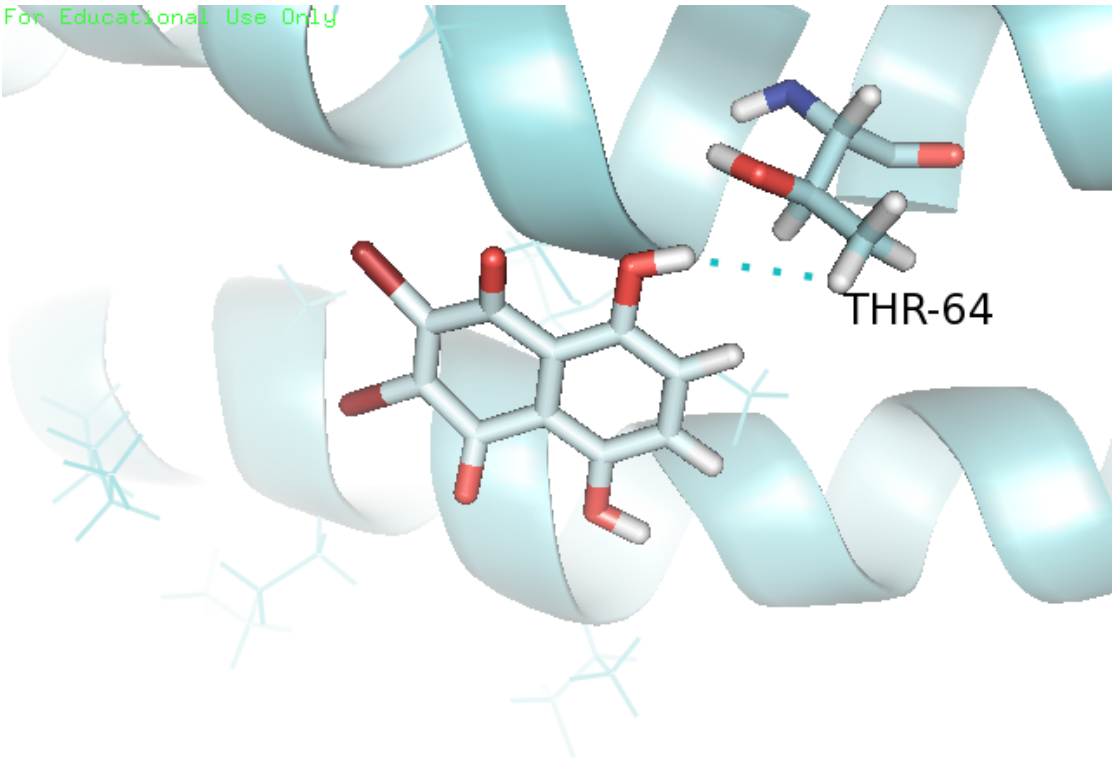


FIGURE 4.2: Compound 2,3-dibromo-5,8-dihydroxy naphthalene-1,4-dione that shows Binding affinity with ATP synthase subunit C

It is involved in the adenosine ribonucleotides *de novo* biosynthesis pathway.

Table 4.2 is showing all the 10 top selected compounds with their binding affinities and interactions for ATP synthase subunit C. Figure 4.2 is showing the interaction of compound 2, 3-dibromo-5,8-dihydroxy naphthalene-1,4-dione that have the lowest binding affinity and have interaction with THR64. THR64 shows interaction with the Hydroquinone moiety as compared to the other parts of the compounds. Hydroquinone moiety of compound 2, 3-dibromo-5,8-dihydroxy naphthalene-1,4-dione possesses this interaction.

4.2.2 Chromosomal Replication Initiator Protein DnaA

It is a cytoplasmic protein that is synthesized by the dnaA gene. It is involved in ATP binding, DNA replication origin binding, initiation and also regulation. It is involved in the two-component system pathway. The molecular weight of chromosomal replication initiator protein DnaA is 5.13 kDa.

The top interacted compounds with the highest affinities with the chromosomal replication initiator protein DnaA are shown in Table 4.3 and Figure 4.3 is showing the interaction of Compound 3-methyl-1-(1,4,5,8-tetramethoxynaphthalen-2-yl) but-3-en-1-olthat shows the lowest Binding affinity and highest interactions with residues GLN388, LYS423, ARG387 and ARG379 with Chromosomal replication initiator protein DnaA. Moiety 1,4,5,8-tetramethoxynaphthalene of compound 3-methyl-1-(1,4,5,8-tetramethoxynaphthalen-2-yl) but-3-en-1-ol possess all these interactions.

TABLE 4.3: Compound's name, binding affinity and Residue interaction of top 10 compounds with Chromosomal replication initiator protein DnaA

Compound	**S-Score/Binding Affinity***	**Residues interaction****	**No of interactions*****
3-methyl-1-(1,4,5,8-tetramethoxynaphthalen-2-yl) but-3-en-1-ol	-12.4461	GLN388 LYS423 ARG387 ARG379	4
1,8-dihydroxy-3-methoxy-6-methylanthracene-9,10-dione	-10.7430	LYS 423 GLN 388 ARG379	3
2,5,8-trihydroxynaphthalene-1,4-dione	-10.5229	LYS 307 LYS 404	2
(E)-4-(1,4,5,8-tetramethoxynaphthalen-2-yl)but-3-en-2-one	-9.5908	ARG379 LYS423 GLN388 ARG387	4
2-hydroxynaphthalene-1,4-dione	-9.3861	LYS404 LYS307	2
2-bromo-1,4,5,8-tetramethoxy-naphthalene	-9.2806	LYS 423 GLN 388	2
2-bromo-5-hydroxynaphth-alene-1,4-dione	-9.1409	LYS 307 LYS 404	2

2-bromo-5,8-dihydroxy-naphthalene-1,4-dione	-9.1408	LYS 307 LYS 404	2
2-(2,5-dimethoxy-4-methylbenzoyl)-3-methoxybenzoic acid	-9.0264	ARG 387 ARG 379 THR 416	3
2,3-dibromo-5,8-dihydroxynaphth-alene-1,4-dione	-8.8189	ARG379 LYS423	2

*The intensity of the binding relationship between a protein and compound (ligand)

** Residues that are involved in the interaction

*** No of interactions formed between active site and ligand

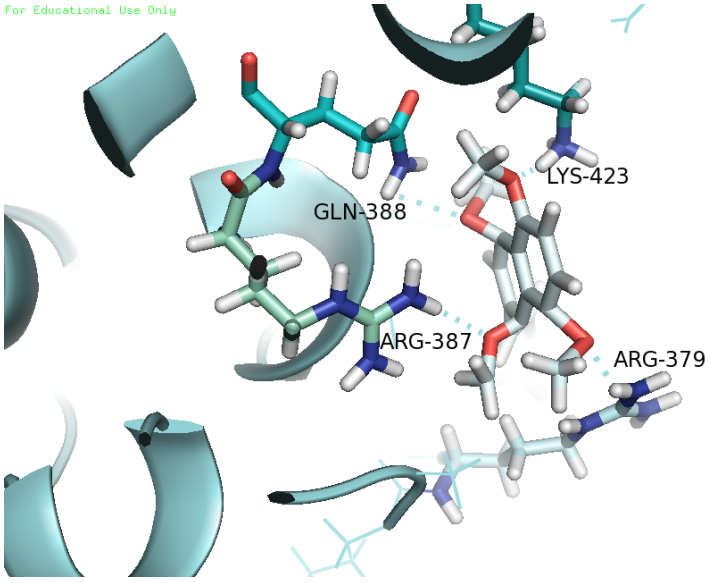


FIGURE 4.3: Compound 3-methyl-1-(1,4,5,8-tetramethoxynaphthalen-2-yl) but -3-en-1- ol that shows Binding affinity with Chromosomal replication initiator protein DnaA.

4.2.3 Thil Domain-containing Protein

Thil domain-containing protein is synthesized by CBO0129. Its molecular functions include ATP binding, sulfurtransferase activity, tRNA Adenylyl transferase

activity and tRNA binding. Biologically it is involved in thiamine biosynthetic process, thiamine diphosphate biosynthetic process, thiazole biosynthetic process and tRNA 4-thiouridine biosynthesis. It's a cytoplasmic protein with molecular weight 37.17 kDa. It is involved in Sulfur relay system.

Figure 4.4 is showing the interactions of (E)-4-(4-bromo-2,5-dimethoxyphenyl) -4-oxobut-2-enoic acid with Thil domain-containing protein by interacting with MET7, SER9 and GLY107 with binding affinity -10.8301. (E)-4-oxobut-2-enoic acid moiety of this compound possess all these interactions of Thil domain-containing protein.

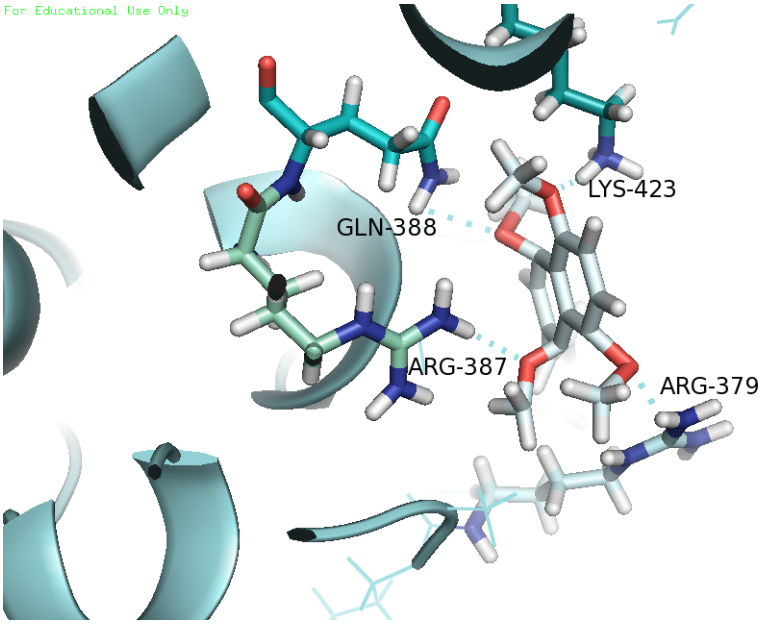


FIGURE 4.4: Compound (E)-4-(4-bromo-2, 5-dimethoxyphenyl) -4-oxobut-2-enoic acid that shows residue interaction with Thil domain-containing protein.

Table 4.4 is showing the top 10 docked compounds along with their binding affinities and interaction.

TABLE 4.4: Compounds name, binding affinity and Residue interaction of top 10 compounds with Thil domain-containing protein

Compound	**S-Score/binding affinity***	**Residues interactions****	**No of interactions*****
Lapachol	-11.7861	SER9	1

(E)-4-(4-bromo-2,5-dimethoxyphenyl)-4-oxobut-2-enoic acid	-10.8301	MET7 SER9 GLY107	3
5-(heptadec-1-en-2-yloxy)naphthalene-1,4-dione	-10.5999	GLY188	1
2-(2,5-dimethoxy-4-methylbenzoyl)-3-methoxybenzoic acid	-10.2562	SER9 GLY107	2
(E)-4-(1,4,5,8-tetramethoxynaphthalen-2-yl)but-3-en-2-one	-9.9548	SER9	1
1,8-dihydroxy-3-methoxy-6-methylanthracene-9,10-dione	-9.7721	ASP13 ASN39 SER185	3
4-(2,5-dimethoxyphenyl)-2-ethoxybut-3-yn-2-yl acetate	-9.7459	SER9 GLY107	2
4-(2,5-dimethoxyphenyl)-2-hydroxybut-3-yn-2-yl acetate	-9.4933	Ser9 Gly188	2
1,4,5,8-tetramethoxy-2-naphthoic acid	-9.3773	Asp13	1
2-bromo-3-pentylnaphtha-lene-1,4-dione	-9.2856	Ser9	1

*The intensity of the binding relationship between a protein and compound (ligand)

** Residues that are involved in the interaction

*** No of interactions formed between active site and ligand

4.2.4 Putative guanosine 3',5' -bis- pyrophosphate (PpGpp)

Putative guanosine 3',5'-bis-pyrophosphate (PpGpp) synthesis/degradation protein is synthesized by CBO0185. Its molecular functions include GTP Di phosphokinases activity and biologically it is involved guanosine tetraphosphate metabolic process. Its molecular weight is 30.88 KDa and it takes part in the purine metabolic pathway. The top 10 compounds with their binding affinities and interactions with PpGpp synthesis/degradation protein are shown in Table 4.5.

Figure 4.5 is showing the interactions of compound 2-(2,5-dimethoxy-4- methylbenzoyl) -3-methoxybenzoic acid that shows the lowest binding energy that is -13.1221 along with residue interaction that includes His156, 2/lys48, Gln142, Arg78 with PpGpp synthesis/degradation protein. His156 and Gln142 interact with 2,5- dimethoxy-4- methylbenzaldehyde moiety while Arg78 and 2/ lys48 interact with 3-methoxybenzoic acid moiety.

TABLE 4.5: Compound's name, binding affinity and Residue interaction of top 10 compounds with Putative guanosine 3',5'-bis-pyrophosphate (PpGpp) synthesis/degradation protein

Compound Name	**S-score/ binding affinity***	**Residues interaction ****	**No of interactions *****
2-(2,5-dimethoxy-4-meth-ylbenzoyl) -3-methoxybenzoic acid	-13.1221	His156 2/lys48 Gln142 Arg78	5
(E)-4-(1,4,5,8-tetra-methoxynaphthalen -2-yl)but-3-en-2-one	-11.8998	Lys56 Lys48 2/ser52	4
2-bromo-1,4,5,8-tetra-methoxynaphthalene	-11.7395	Lys48 2/ser52	3

1,4,5-trimethoxy-2-methylanthracene-9,10-dione	-11.5037	2/ Lys48 Arg78	3
(E)-2-hydroxy-2-methyl-4-(1,4,5,8-tetramethoxynaphthalen-2-yl)but-3-enyl acetate	-10.8476	Lys48 2/ser52	3
1,4,5,8-tetramethoxy-2-naphthoic acid	-10.7450	2/Lys48 2/ser52	5
4-(2,5-dimethoxyphenyl)-2-ethoxybut-3-yn-2-yl acetate	-10.4253	Arg78 Gln142 Arg78 Lys48 His156	4
5,8-dioxo-5,8-dihydronaphthalen-1-yl tetra decanoate	-10.3810	Arg 78 Arg46	2
4-(2,5-dimethoxyphenyl)-2-hydroxybut-3-yn-2-yl acetate	-10.3668	Lys48 Arg78	2
2-bromo-5,8-dihydroxy-naphthalene-1,4-dione	-10.2339	Arg78	1

*The intensity of the binding relationship between a protein and compound (ligand)

** Residues that are involved in the interaction

*** No of interactions formed between active site and ligand

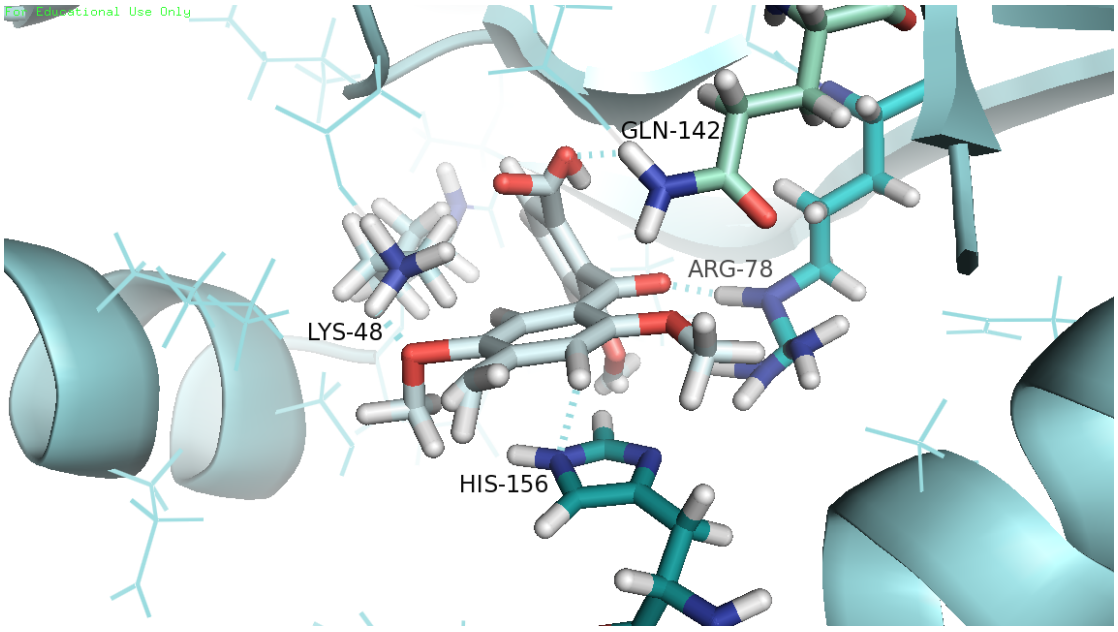


FIGURE 4.5: Compound 2 - (2,5-dimethoxy-4 -methylbenzoyl) -3- methoxybenzoic acid that shows binding energy along with residue interaction with Putative guanosine 3',5'-bis-pyrophosphate (PpGpp) synthesis / degradation protein.

4.2.5 Putative Metal Dependent Phosphohydrolase

Putative metal dependent phosphohydrolase protein is a cytoplasmic protein with molecular weight 2.58 KDa. It performs hydrolase activity. This protein is synthesized by CBO1098. Its biological functions include RNA repair and tRNA 3'-terminal CCA addition. It takes part in a number of pathways that include metabolic pathways in such diverse environments.

The top 10 interacted compounds with Putative metal dependent phosphohydrolase protein are shown in Table 4.6.

Figure 4.6 is showing the interaction of compound 2-(2,5-dimethoxy-4-methylbenzoyl)-3-methoxybenzoic acid that shows highest number of residue interaction that include ASN8, ASP5 and GLU 43 with binding affinity -12.3890.

ASN8 interacts with 1,4-dimethoxy -2- methylbenzenemoiety of compound 2- (2,5-dimethoxy-4-methylbenzoyl) -3-methoxybenzoic acid while ASP5 and GLU43 interacts with the other moiety benzoic acid.

TABLE 4.6: Compound's name, binding affinity and Residue interaction of top 10 compounds with Putative guanosine 3',5'- bis-pyrophosphate (PpGpp) synthesis / degradation protein

Compound Name	**S-score / Binding Affinity***	**Residue Interaction ****	**No of Interactions *****
(E)-2-hydroxy-2-methyl -4-(1,4,5,8-tetramethoxynaphthalen-2-yl)but-3-enyl acetate	-14.6802	Lys48 2/Ser52	3
Lapachol	-12.6511	Arg78 Lys48	2
2-(2,5-dimethoxy-4-methyl benzoyl)-3-methoxybenzoic acid	-12.3890	ASN8 ASP5 GLU43	3
3-methyl-1-(1,4,5,8-tetramethoxynaphthalen-2-yl) but-3-en-1-ol	-12.0151	Tyr44 Asp5	2
2-(4-chlorobenzylthio)-3 -thiomorpholinonaphthalene-1,4-dione	-11.9529	Asp5	1
1,8-dihydroxy-3-methoxy-6-methylanthracene-9,10-dione	-10.9505	Glu43	1
1,4,5-trimethoxy-2-methylanthracene-9, 10-dione	-10.6678	Asp5	1
6,7-dichloro-5,8-dioxo-5,8-dihydronaphthalene -1,4-diyl diacetate	-10.6422	Tyr44	1

2-(4-chlorobenzylthio) -3-(piperidin-1-yl) naphthalene-1,4-dione	-10.3361	Asp5	1
6-bromo-5,8-dioxo-5, 8-dihydronaphthalene -1,4-diyl diacetate	-10.6678	Asp5	1

*The intensity of the binding relationship between a protein and compound (ligand).

** Residues that are involved in the interaction.

*** No of interactions formed between active site and ligand.

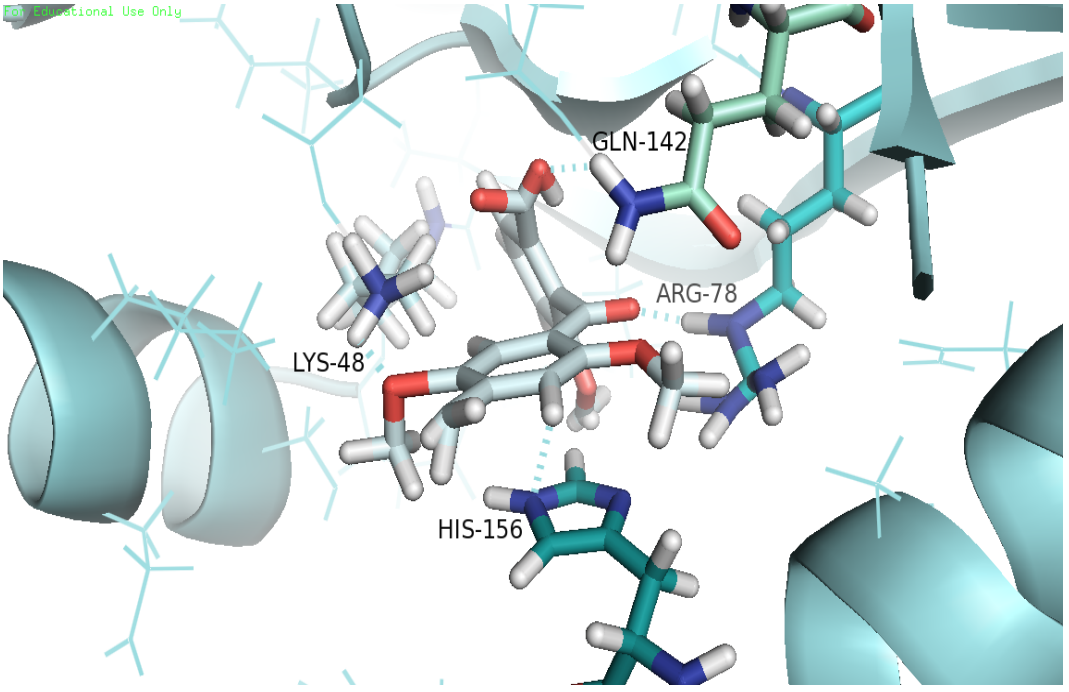


FIGURE 4.6: Compound 2- (2,5-dimethoxy-4-methylbenzoyl) -3- methoxybenzoic acid that shows binding energy along with residue interaction with Putative guanosine 3',5'-bis-pyrophosphate (PpGpp) synthesis/degradation protein.

4.2.6 Putative Metal Dependent Phosphohydrolase

Putative metal dependent phosphohydrolase protein is a cytoplasmic protein with molecular weight 2.58 KDa. It performs hydrolase activity. This protein is synthesized by CBO1098. Its biological functions include RNA repair and tRNA

3'-terminal CCA addition. It takes part in a number of pathways that include metabolic pathways in such diverse environments.

The top 10 interacted compounds with Putative metal dependent phosphohydrolase protein are shown in Table 4.7 and Figure 4.7 is showing the interaction of compound 2-(2,5-dimethoxy-4-methylbenzoyl)-3-methoxybenzoic acid that shows highest number of residue interaction that include ASN8, ASP5 and GLU 43 with binding affinity -12.3890. ASN8 interacts with 1,4-dimethoxy -2 -methyl benzenemoiety of compound 2- (2,5-dimethoxy-4- methylbenzoyl) -3-methoxybenzoic acid while ASP5 and GLU43 interacts with the other moiety benzoic acid.

TABLE 4.7: Compounds name, binding affinity and Residue interaction of top 10 compounds with Putative metal-dependent phosphohydrolase protein

Compound Name	**S-score / Binding Affinity***	**Residue Interaction ******	**No of Interactions *******
(E)-2-hydroxy-2-methyl -4-(1,4,5,8-tetramethoxynaphthalen-2-yl) but-3-enyl acetate	-14.6802	Lys48 2/Ser52	3
Lapachol	-12.6511	Arg78 Lys48	2
2-(2,5-dimethoxy-4-methylbenzoyl)-3-methoxybenzoic acid	-12.3890	ASN8 ASP5 GLU43	3
3-methyl-1-(1,4,5,8 -tetramethoxynaphthalen-2-yl)but-3-en-1-ol	-12.0151	Tyr44 Asp5	2
2-(4-chlorobenzylthio) -3-thiomorpholinonaphthalene-1,4-dione	-11.9529	Asp5	1

1,8-dihydroxy-3-methoxy-6-methylanthracene-9,10-dione	-10.9505	Glu43	1
1,4,5-trimethoxy-2-methylanthracene-9,10-dione	-10.6678	Asp5	1
6,7-dichloro-5,8-dioxo-5,8-dihydro naphthalene-1, 4-diyl diacetate	-10.6422	Tyr44	1
2-(4-chlorobenzyl thio)-3-(piperidin-1-yl)naphthalene-1,4-dione	-10.3361	Asp5	1
6-bromo-5,8-dioxo-5,8-dihydr-onaphthalene-1, 4-diyl diacetate	-10.6678	Asp5	1

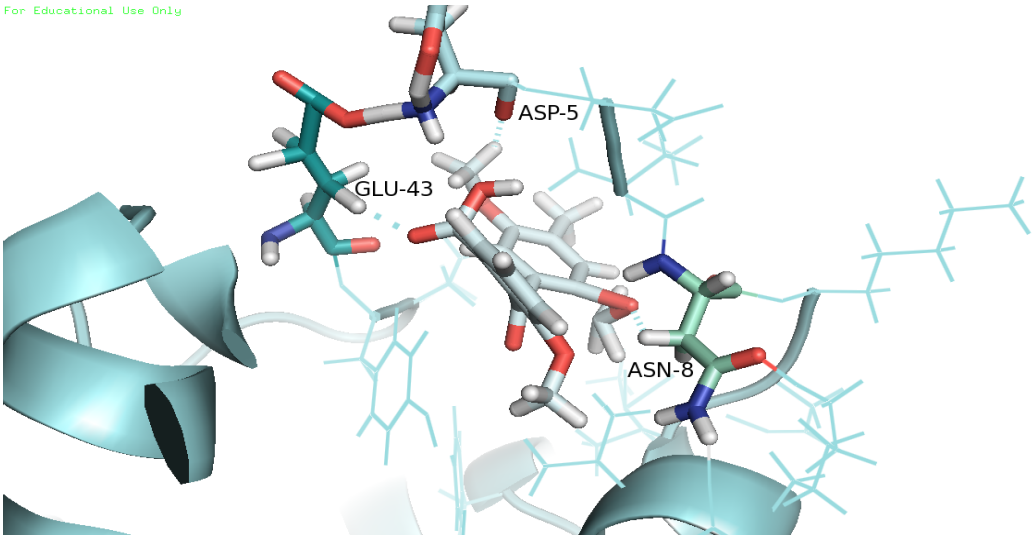


FIGURE 4.7: Compound 2-(2, 5 -dimethoxy-4 - methylbenzoyl) -3 - methoxybenzoic acid that shows number of residue interaction with Putative metal dependent phosphohydrolase protein.

4.2.7 UDP-N-acetylmuramoyl-tripeptide-D-alanyl-D- alanine Ligase

UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D- alanine ligase is synthesized by murF with molecular weight 51.90 kDa. It is molecularly involved in ATP binding, UDP-N- acetylmuramoylalanyl-D-glutamate-2,6-diaminopimelate ligase activity.

Biologically it is involved in different processes of the cell which may include and involve cell cycle, division, cell wall organization, peptidoglycan biosynthetic process and cell shape regulation. It is involved in many pathways that include vancomycin resistance, peptidoglycan biosynthesis, some metabolic pathways and biosynthesis.

The interaction, binding affinities and interactions of top 10 compounds with UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D-alanine ligase are shown in Table 4.8.

Figure 4.8 is showing the interactions of compound 1,4,5,8-tetramethoxy-2-naphthoic acid that shows lowest Binding affinity with UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D-alanine ligase. It shows interaction with Thr320 that interacts with moiety 1,4,5,8-tetramethoxynaphthalene.

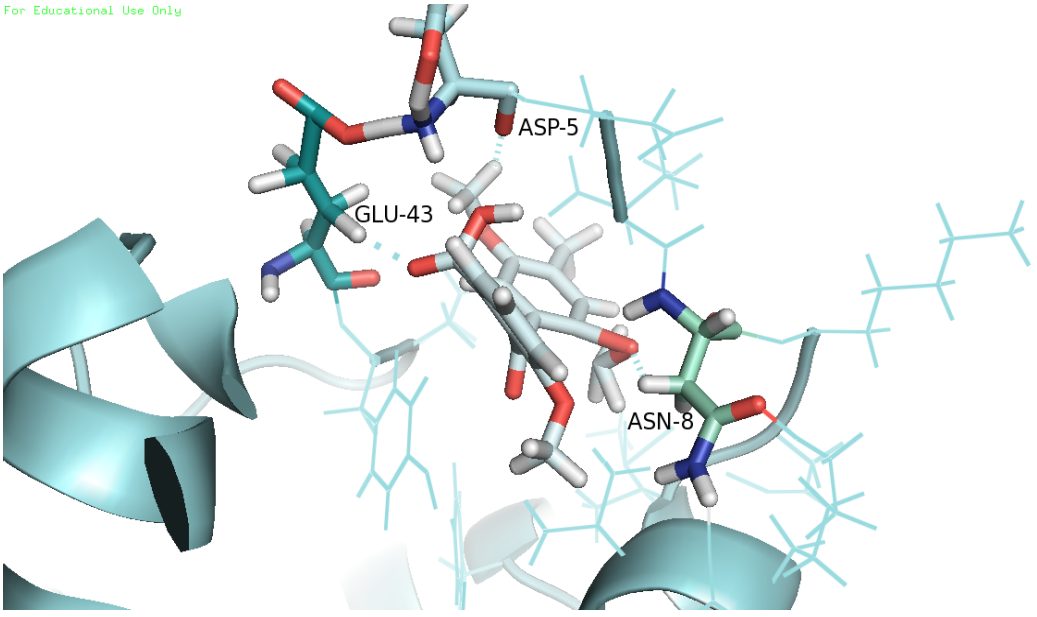


FIGURE 4.8: Compound 1,4,5,8-tetramethoxy-2-naphthoic acid that shows highest Binding affinity with UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D-alanine ligase.

TABLE 4.8: Compounds name, binding affinity and Residue interaction of top 10 compounds with UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D-alanine ligase

Compound Name	**S-score / Binding Affinity***	**Residue Interaction ****	**No of Interactions *****
1,4,5,8-tetramethoxy-2-naphthoic acid	-12.9904	2/Thr320	2
2-bromo-1,4,5,8-tetramethoxynaphthalene	-11.6365	2/Thr320	2
(E)-2-hydroxy-2-methyl-4-(1,4,5,8-tetramethoxynaphthalen-2-yl) but-3-enyl acetate	-11.2598	Thr320	1
1,4,5,8-tetramethoxynaphthalene	-11.0901	2/Thr320	1
2-(3-ethyl-1,4-dioxo-1,4-dihydronaphthalen-2-yl)acetic acid	-10.8672	Asn289	1
4-(2-methylbut-3-yn-2-yloxy)phenol	-10.8069	Asp124	1
2-bromo-5,8-dihydroxynaphthalene-1, 4-dione	-10.8069	Thr320	1
2-bromo-5-hydroxynaphthalene-1,4-dione	-10.7980	Thr320	1
Lapachol	-10.7054	Thr320	1

2,5,8-trihydroxyn-aphthalene-1, 4-dione	-10.6042	Asn289	1

*The intensity of the binding relationship between a protein and compound (ligand).

** Residues that are involved in the interaction.

*** No of interactions formed between active site and ligand.

4.2.8 Chemotaxis

Chemotaxis protein is of 33.60 kDa molecular weight and is synthesized by CheV gene. It is a cytoplasmic protein that is involved in DNA binding and biologically involved in phosphorelay signal transduction system and transcription regulation and DNA-templated. It takes part in 2 pathways in Two-component system and bacterial chemotaxis.

The compounds, their binding affinities and residue interactions with the chemotaxis are shown in Table 4.9. Figure 4.9 is showing the GLY 126, LEU 137 and GLN74 interaction of compound 2,3-dibromo-5,8-dihydroxy naphthalene-1,4-dione with chemotaxis protein. Residue GLY 126 and LEU 137 interacts with hydroquinone moiety while GLN74 interacts with 2,3-dibromocyclohex-2-ene-1,4-dionemoiety of compound 2,3-dibromo-5,8-dihydroxynaphthalene-1,4-dione.

TABLE 4.9: Compounds name, binding affinity and Residue interaction of top 10 compounds with Chemotaxis protein

Compound Name	**S-score / Binding Affinity***	**Residue Interaction ******	**No of Interactions *******
2,3-dibromo-5,8-dihydroxynapht-halene-1,4-dione	-10.5992	GLY 126 LEU 137 GLN74	3

2-chloro-5,8-dihydroxynaphthalene-1,4-dione	-9.8197	LEU 137 GLN 74	2
6,7-dichloro-5,8-dimethoxynaphthalene-1,4-dione	-9.8036	GLN 74	1
5,8-dihydroxy-2,3-dihydronaphthalene-1,4-dione	-9.6453	LEU 137 GLN 74	2
4-(2,5-dimethoxyphenyl)but-3-yne-1,2-diol	-8.7919	LEU137 TYR 26	2
Juglone	-8.7648	LEU 137 GLY 126 GLN74	3
2-(4-chlorobenzylthio)-3-morpholinonaphthalene-1,4-dione	-8.6547	GLN 74	1
2-hexylnaphthalene-1,4-dione	-8.6206	GLN 74	1
2-bromo-3-pentylnaphthalene-1,4-dione	-8.5125	GLN 72	1
2-bromo-7-hydroxynaphthalene-1,4-dione	-8.3657	GLN 74 GLY126	2

*The intensity of the binding relationship between a protein and compound (ligand).

** Residues that are involved in the interaction.

*** No of interactions formed between active site and ligand.

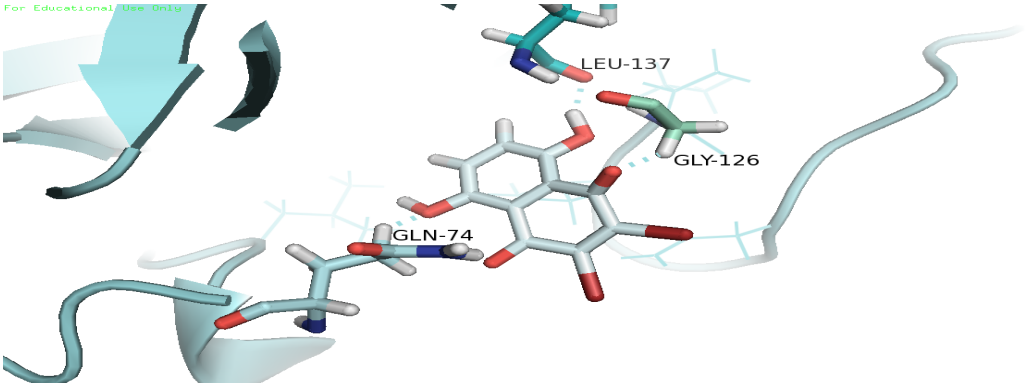


FIGURE 4.9: Compound 2, 3-dibromo-5, 8-dihydroxynaphthalene-1, 4-dione that shows affinity with chemotaxis protein.

4.3 Epitope-Based Vaccine Identification

4.3.1 Epitope Target Identification

2 proteins named as a pre-protein translocase, SecG subunit and UDP -N- acetyl-muramyl tripeptide synthetase-like protein were selected based on cellular localization e.g., extracellular membrane proteins. Both these proteins are non-homologous and essential proteins.

4.3.2 Structure Analysis

The physiochemical properties computed via protaparm of preprotein translocase, SecG subunit shows that it has 77 amino acids with 8333.94 kDa. The nature of protein is positive that as indicated by its theoretical isoelectric protein (PI) is 10.02. PI above 7 shows the positive nature of protein while below 7 shows the negative nature of the protein. The computed physical properties of UDP-N-acetylmuramyl tripeptide synthetase-like protein shows that it has 453 amino acids with 51525.55 kDa. It also has a positive nature because its PI is 8.76. Table 4.8 is showing all physicochemical properties that include stability index, instability

index, aliphatic index etc. of both proteins. According to the stability index of both proteins are considered stable proteins.

DIANNA 1.1 shows that there are no disulfides bonds are present in preprotein translocase, SecG subunit while MUR ligase family protein shows 31 disulfide bonds. Table 4.9 is showing the bonds for MUR ligase family protein along with distance and bond scores. Vaxijen shows that preprotein translocase, SecG subunit is a probable antigen with a score of 0.5628 while MUR ligase family protein is non-antigen with a score of 0.3451. According to AllerTop results, both proteins are non-allergen.

PSIPRED was used for the analysis of the secondary structure of both proteins. Figure 4.10 and Figure 4.11 are showing the conformations of the secondary structure of both proteins. To check the transmembrane topology of both proteins TMHMM was used. TMHMM results of preprotein translocase, SecG subunit shows that its 1 to 3 amino acids and 76 to 77 were exposed to the outside surface, while 22 to 52 were embedded in the inside surface and 4 to 21 and 53 to 75 were present in the transmembrane helix. TMHMM results for MUR ligase family protein shows that its 1 to 453 amino acids were exposed to the outside surface.

TABLE 4.10: Physiochemical parameters of both proteins computed via Protparam

Parameters	**Preprotein translocase, SecG subunit**	**UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein)**
Mol. Weight	8333.94 kDa	51525.55 kDa
No. of amino acids	77	453
Theoretical pI	10.02	8.76
Instability index (II)	35.19	26.48
No. of Negatively Charged Residues (Asp + Glu)	2	52

No. of Positively Charged Residues (Arg + Lys)	6	60
Aliphatic Index	127.66	96.60
Grand average of Hydropathicity (GRAVY)	0.813	-0.200
Atomic Composition	Carbon C 376 Hydrogen H 628 Nitrogen N 96 Oxygen O 107 Sulfur S 4	Carbon C 2340 Hydrogen H 3694 Nitrogen N 586 Oxygen O 686 Sulfur S 16
Half-life	30 hours (mammalian reticulocytes, in vitro). Greater than 20 hours (yeast, in vivo) Greater than 10 hours (Escherichia coli, in vivo).	30 hours (mammalian reticulocytes, in vitro). Greater than 20 hours (yeast, in vivo) Greater than 10 hours (Escherichia coli, in vivo)

TABLE 4.11: Disulfide bonds along with distance, bond and bond score for UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein)

Cysteine Sequence Position	**Distance**	**Bond**	**Score**
2 - 97	95	XXXXMCKINIK-PGIVACFVENF	0.01059
2 - 212	210	XXXXMCKINIK-ADSKFCKKCKH	0.0107
2 - 215	213	XXXXMCKINIK-KFCKKCKHPYS	0.01086
2 - 234	232	XXXXMCKINIK-LGSYYCENCGY	0.01055

2 - 237	235	XXXXMCKINIK-YYCENCGYKRP	0.01126
2 - 283	281	XXXXMCKINIK-VYNALCAYSMA	0.01116
2 - 422	420	XXXXMCKINIK-EDIKSCREETV	0.01079
97 - 212	115	PGIVACFVENF-ADSKFCKKCKH	0.01569
97 - 215	118	PGIVACFVENF-KFCKKCKHPYS	0.02294
97 - 234	137	PGIVACFVENF-LGSYYCENCGY	0.01046
97 - 237	140	PGIVACFVENF-YYCENCGYKRP	0.03147
97 - 283	186	PGIVACFVENF-VYNALCAYSMA	0.01055
97 - 422	325	PGIVACFVENF-EDIKSCREETV	0.082
212 - 215	3	ADSKFCKKCKH-KFCKKCKHPYS	0.04107
212 - 234	22	ADSKFCKKCKH-LGSYYCENCGY	0.02156
212 - 237	25	ADSKFCKKCKH-YYCENCGYKRP	0.037
212 - 283	71	ADSKFCKKCKH-VYNALCAYSMA	0.01037
212 - 422	210	ADSKFCKKCKH-EDIKSCREETV	0.01041
215 - 234	19	KFCKKCKHPYS-LGSYYCENCGY	0.03006

215 - 237	22	KFCKKCKHPYS-YYCENCGYKRP	0.04977
215 - 283	68	KFCKKCKHPYS-VYNALCAYSMA	0.01038
215 - 422	207	KFCKKCKHPYS-EDIKSCREETV	0.01043
234 - 237	3	LGSYYCENCGY-YYCENCGYKRP	0.0117
234 - 283	49	LGSYYCENCGY-VYNALCAYSMA	0.01052
234 - 422	188	LGSYYCENCGY-EDIKSCREETV	0.02373
237 - 283	46	YYCENCGYKRP-VYNALCAYSMA	0.01052
237 - 422	185	YYCENCGYKRP-EDIKSCREETV	0.01748
283 - 422	139	VYNALCAYSMA-EDIKSCREETV	0.01593

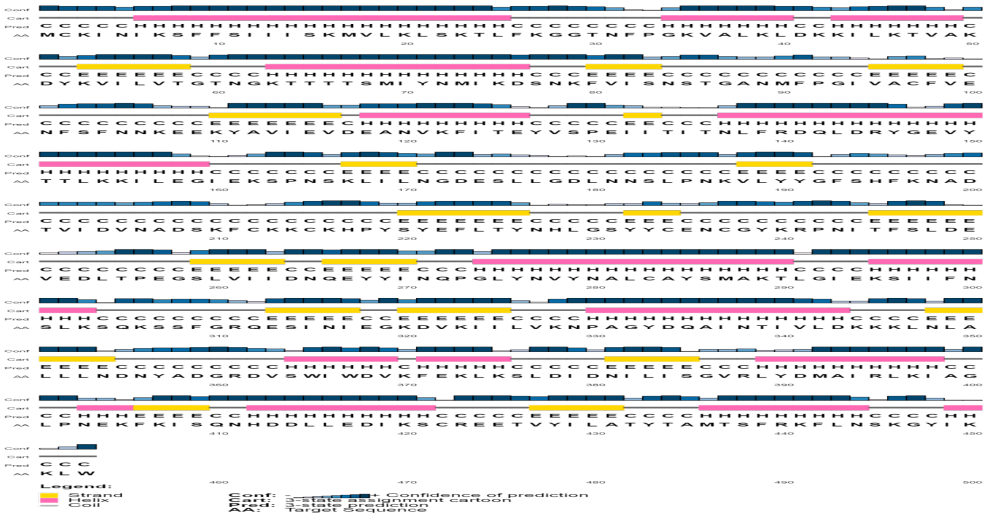


FIGURE 4.10: PSPIRED analysis of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein)

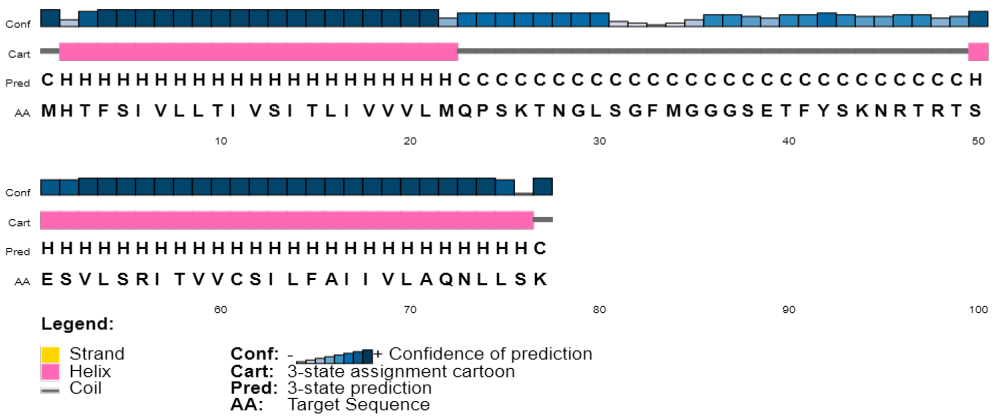


FIGURE 4.11: PSPIRED analysis of Preprotein translocase, SecG subunit B-cell epitope recognition

There are different features of potential B-cell epitopes that directs them to recognize and then activate the defense responses. Preprotein translocase, SecG subunit has only one B-cell epitope of 14 residues with 75% specificity (Table 4.11). UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) shows several predicted epitopes but only 6 epitopes were selected that have the highest antigenicity score (only antigenic epitopes) that was computed through Vaxijen 2.0 (4.12).

Bepipred linear epitope results of Preprotein translocase, SecG subunit predicts 2 peptide epitopes at threshold 0.350 with average -0.789, minimum -0,096 and maximum 1.205 values. The first peptide SKTNGL starts from position 25 and ends at 30 with peptide length 6 while the second peptide FMGGGSETFYSKN-RTRTSE starts from position 33 and ends at position 51 with 19 peptide lengths Figure 4.12 A. Karplus & Schulz flexibility results predict average values as 0.984, minimum at 0.892 and maximum value lies at 1.108 Figure 4.12 B. Chou & Fasman Beta-Turn results predict the average value at 0.904, while the minimum at 0.536 and maximum at 1.289 at threshold 0.904 Fig 4.12 C. An Emini Surface Accessibility result predicts average value 1.000, while minimum 0.050 and maximum 6.639 at threshold 1 with two peptides; first peptide QPSKTN starting from position 23 and ends at 28 and second peptide ETFYSKNRTRTS start from position 39 and ends at 50 Fig 4.12 D. Kolaskar & Tongaonkar Antigenicity result

predicts two peptides; SIVLLTIVSITLIVVVLM starting from position 5 to 22 and LSRITVVCSILFAIIVLAQN starting from position 54 to 73 with average values 1.071 while the minimum at 0.886 and maximum at 1.244 at threshold 1.071 Fig 4.12 E. Parker Hydrophilicity result predicts average value at -0.165, while the minimum at -6.457 and maximum at 5.729 at threshold -0.16 Fig 4.12 F.

Bepipred Linear Epitope results for UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) predicts the average values -0.226, minimum at -0.005 and maximum at 1.656 with threshold 0.350 Figure 4.13 A. Karplus & Schulz Flexibility results give average values at 0.998, minimum at 0.900 and maximum at 1.123 at threshold 0.998 4.13 B. Chou & Fasman Beta-Turn results predict average value at 0.982, minimum at 0.650 and maximum at 1.343 with a threshold of 0.982 Fig 4.13 C. Emini Surface Accessibility result predicts 11 peptides starting "TNGKTTT (60-66), MIKDSN (73-78), FNNKEEK (104-110), DQLDRYGE (141-148), IEKSPN (160-165), CKHPYS (215-220), CGYKRP (237-242), NQEYYIN (265-271), LKSQKSSFGRQE (302-313), NDNYAD (254-359) and PNEKFKISQNH (402-412)" with average value 1, minimum 0.042 and maximum 7.210 at threshold 1 Fig 4.13 D. Kolaskar & Tongaonkar Antigenicity predicts 17 peptides that include "FFSIIISKMVLKLSKT (9-24), KVALKLDKK (35-43), LKTVAKDYKVILVT (45-58), GIVACFVEN (93-101), YAVIEV (111-116), ITEYVSPEI (124-132), EVYTTLKKI (148-156), NKVLYYGFS (186-194), CKKCKHPYSYEFL (212-224), LGSYYCENC (229-237), EGSLVIID (257-264), LYNVYNALCAYS (275-286), VKIILVKNP (322-330), NTIVL DK (338-344), NL ALLLN (348-354), ISGVRLY (384-390) and VYILAT (427-432)" with average value 1.028, minimum 0.883 and maximum 1.194 with threshold 1.028 Fig 4.13 E. Parker Hydrophilicity results predicted average value 1.143, -5.271 as a minimum and maximum 5.900 at threshold 1.143 Fig 4.13 F.

Discotope 2.0 server was used to predict the discontinuous peptide position with residue name, several contacts, propensity and DiscoTope score. These epitopes are shown on the 3D structure of each protein through pymole (Table 4.13 and 4.14, Figure 4.14 A and Figure 4.14 B). Position of all the predicted B-cell epitopes of both proteins was confirmed by PepSurf is shown in Fig 4.15 A & Fig 4.15 B.

TABLE 4.12: B-cell epitopes present on the surface of Preprotein translocase, SecG subunit predicted via BCPRED along with their starting position and antigenicity scores.

Position	**Epitope**	**Score**	**Vaxijen 2.0 antigenicity score**
32	GFMGGGSETFYSKN	0.992	0.5022

TABLE 4.13: B-cell epitopes present on the surface of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) predicted via BCPRED along with their starting position and antigenicity scores

Position	**Epitope**	**Score**	**Vaxijen 2.0 antigenicity score**
58	TGTNGKTTTTSMIY	0.973	2.0117
99	VENFSFNNKEEKYA	0.905	0.7097
327	VKNPAGYDQAINTI	0.876	0.4101
83	SNSTGANMFPGIVA	0.863	0.8173
115	EVDEANVKFITEYV	0.849	0.4357
354	NDNYADGRDVSWIW	0.722	1.0002

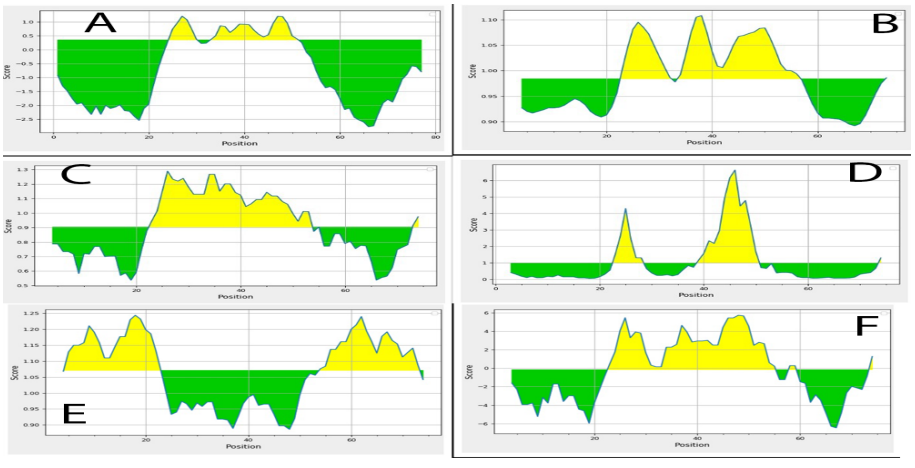


FIGURE 4.12: Preprotein translocase, SecG subunit a Bepipred Linear Epitope prediction results, b flexibility analyses using Karplus and Schulz flexibility scale, c beta turns analyses in structural polyprotein using Chou and Fasman beta-turn prediction, d surface accessibility analyses using Emini surface accessibility scale, e prediction of antigenic determinants using Kolaskar and Tongaonkar antigenicity scale, f hydrophilicity prediction using Parker hydrophilicity

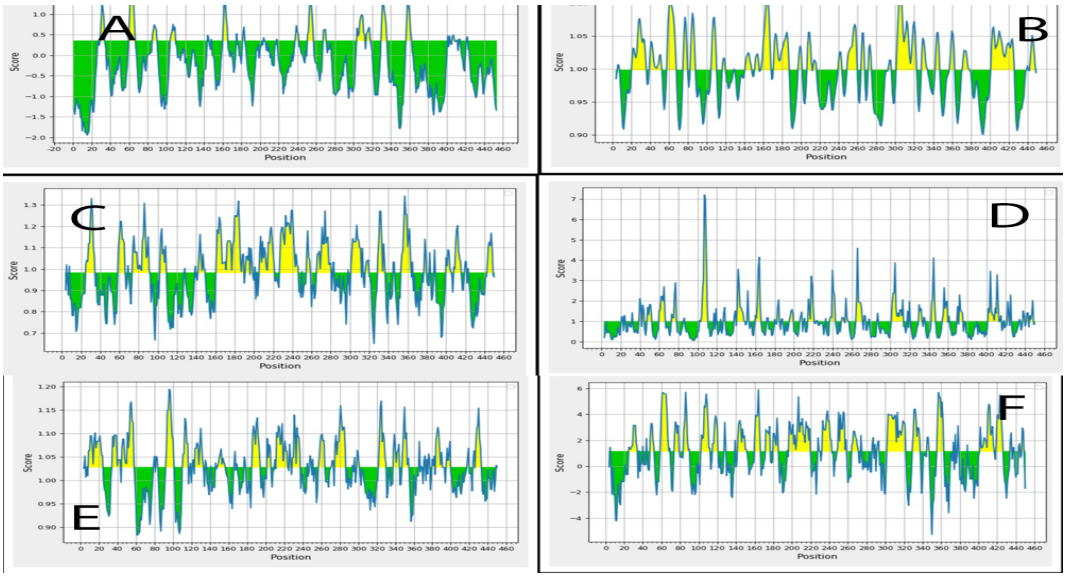


FIGURE 4.13: UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) a Bepipred Linear Epitope prediction results, b flexibility analyses using Karplus and Schulz flexibility scale, c beta turns analyses in structural polyprotein using Chou and Fasman beta-turn prediction, d surface accessibility analyses using Emini surface accessibility scale, e prediction of antigenic determinants using Kolaskar and Tongaonkar antigenicity scale, f hydrophilicity prediction using Parker hydrophilicity

TABLE 4.14: Discontinuous epitopes of Preprotein translocase, SecG subunit predicted through DICOTOPE 2.0 server

Position	**Residue Name**	**No.oF Contact**	**Propensity Score**	**Discotope Score**
26	LYS	12	-1.109	-2.362
27	THR	15	0.577	-1.215
28	ASN	5	1.357	0.626
29	GLY	7	0.929	0.017
30	LEU	3	-1.368	-1.556
31	SER	5	-1.841	-2.205
32	GLY	2	-0.766	-0.908
33	PHE	8	-0.207	-1.103
34	MET	11	0.522	-0.803
35	GLY	0	0.812	0.718
36	GLY	4	0.949	0.380
37	GLY	11	0.944	-0.430

38	SER	14	2.253	0.384
39	GLU	2	2.343	1.843
40	THR	3	2.104	1.517
41	PHE	8	2.078	0.919
42	TYR	7	1.859	0.840
43	SER	0	2.044	1.809
44	LYS	5	3.322	2.365
45	ASN	11	2.104	0.597
46	ARG	13	-0.137	-1.617
47	THR	6	-1.448	-1.971
48	ARG	11	-2.726	-3.678
49	THR	0	-3.897	-3.449

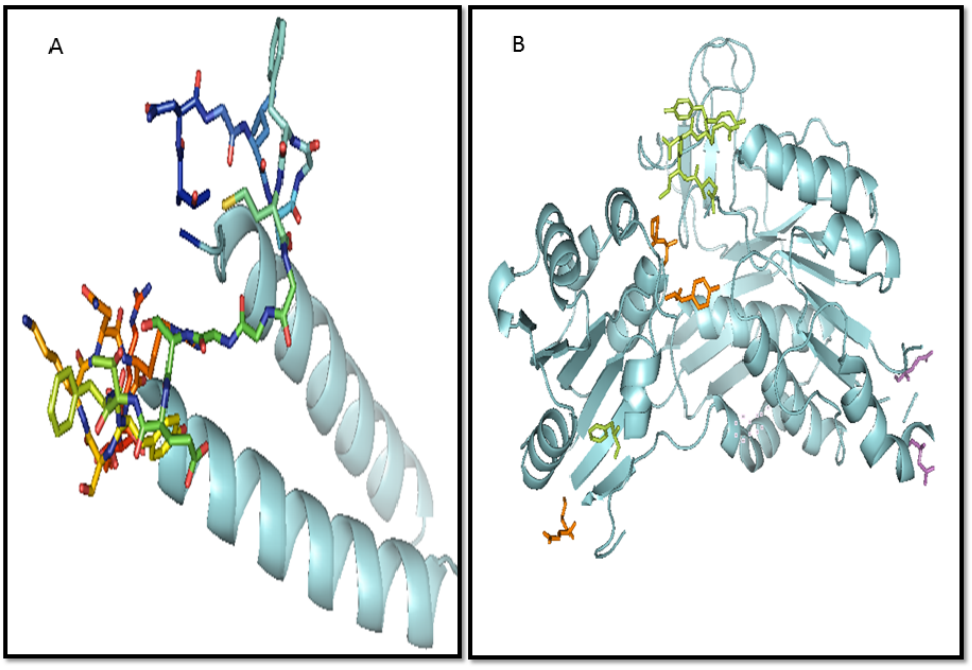


FIGURE 4.14: Site of discontinuous epitopes predicted through Disctope on the 3D structure of **(A)** Preprotein translocase, SecG subunit and **(B)** UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) and presented by sticks through Pymole

TABLE 4.15: Discontinuous epitopes of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) predicted through DICOTOPE 2.0 server

Position	**Residue Name**	**No. of Contact**	**Propensity Score**	**Discotope Score**
108	GLU	0	-3.905	-3.456
105	ASN	4	-2.927	-3.050
142	GLN	9	0.339	-0.735
143	LEU	17	-1.956	-3.686
144	ASP	5	2.429	1.575
145	ARG	9	0.946	-0.198
146	TYR	6	-1.462	-1.984
147	GLY	0	-3.015	-2.668
318	GLU	10	-2.713	-3.551
319	GLY	8	-1.417	-2.174
357	TYR	0	-2.606	-2.306

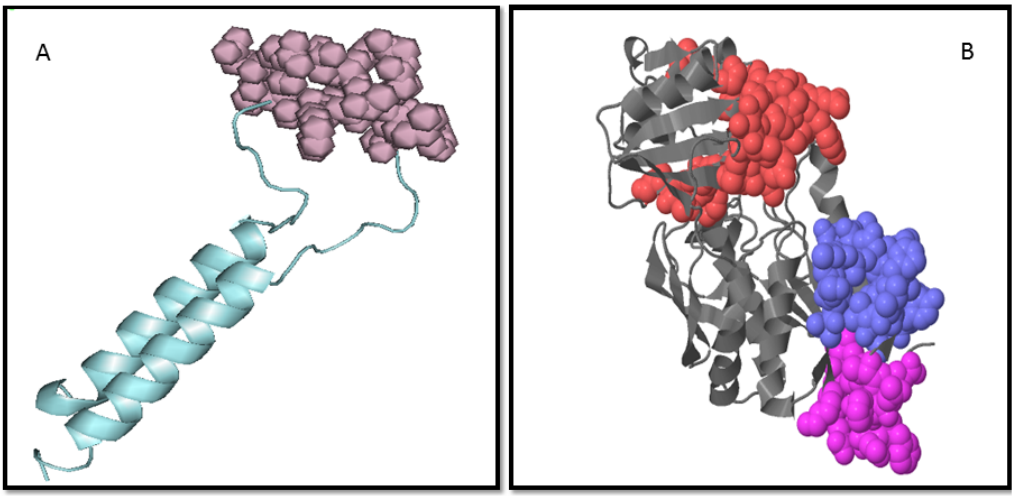


FIGURE 4.15: Site of linear epitopes predicted through PepSurf on the 3D structure of **(A)** Preprotein translocase, SecG subunit and **(B)** UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR ligase family protein) presented by balls

4.3.3 T-cell Epitope Recognition

T-cell epitopes were predicted through online tools known as propred and propred-1 for both proteins. Propred-1 gives the MHC class-I peptides and for MHC class-II peptides propred was used. These tools use matrix base approaches to scan and predict the peptides.

The input of both tools is FASTA sequence of protein at threshold 4% with both proteasome filter and immune proteasome filter mode on. Propred-1 gives three peptides for each protein that shows the presence of 4 MHC class-I alleles (Table 4.15 and 4.16).

Propred gives 28 peptides for UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN) and 19 for the Preprotein translocase, SecG subunit (Table 4.17 and 5.2). Antigenicity scores for all the peptides were checked through Vaxijen 2.0 server. Based on the antigenicity scores, all the peptides were antigenically used for further analysis. Peptide IVLLTIVSI of Preprotein translocase, SecG subunit shows the presence of highest number of alleles i.e. 47 that are

DRB1-0101, "DRB1-0102, DRB1-0301, DRB1-0305, DRB1-0306, DRB1-0307, DRB1-0308, DRB1-0309, DRB1-0311, DRB1-0401, DRB1-0404, DRB1-0405, DRB1-0408, DRB1-0410, DRB1-0421, DRB1-0423, DRB1-0426, DRB1-0701, DRB1-0703, DRB1-0801, DRB1-0802, DRB1-0804, DRB1-0806, DRB1-0813, DRB1-0817, DRB1-1101, DRB1-1102, DRB1-1104, DRB1-1106, DRB1-1107, DRB1-1114, DRB1-1120, DRB1-1121, DRB1-1128, DRB1-1301, DRB1-1302, DRB1-1304, DRB1-1305, DRB1-1307, DRB1-1311, DRB1-1321, DRB1-1322, DRB1-1323, DRB1-1327, DRB1-1328, DRB1-1502, DRB1-1506"

with antigenicity score 0.4421. For UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN) peptide LKLSKTLFK shows the presence of highest number of alleles i.e. 30 that are

"DRB1-0301, DRB1-0305, DRB1-0306, DRB1-0307, DRB1-0308, DRB1-0311, DRB1-0401, DRB1-0402, DRB1-0404, DRB1-0408, DRB1-0410, DRB1-0421, DRB1-0423,

DRB1-0426, DRB1-0804, DRB1-1101, DRB1-1102, DRB1-1104, DRB1-1106, DRB1-1107, DRB1-1114, DRB1-1121, DRB1-1128, DRB1-1305, DRB1-1307, DRB1-1311, DRB1-1322, DRB1-1323, DRB5-0101, DRB5-0105" with antigenic score 0.2046.

TABLE 4.16: MHC class-I allele binding peptides of pre-protein translocase, SecG subunit predicted via Propred-1 with their antigenicity scores

Peptide sequence	**Position**	**MHC class I alleles**	**Vaxijen score**
MHTFSIVLL	0	HLA-A1, HLA-A2, HLA-A*0201,HLA-A*0205	0.8768
HTFSIVLLT	1	HLA-A1, HLA-A2, HLA-A*0201,HLA-A*0205	0.5881
TFSIVLLTI	2	HLA-A1, HLA-A2, HLA-A*0201,HLA-A*0205	1.2617

TABLE 4.17: MHC class-I allele binding peptides of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN) predicted via Propred-1 with their antigenicity scores

Peptide sequence	**Position**	**MHC class I alleles**	**Vaxijen score**
MCKINIKSF	0	HLA-A1, HLA-A2, HLA-A*0201, HLA-A*0205	0.9817
CKINIKSFF	1	HLA-A1, HLA-A2, HLA-A*0201, HLA-A*0205	0.8984
KINIKSFFS	2	HLA-A1, HLA-A2, HLA-A*0201, HLA-A*0205	0.7945

4.3.4 Imperative Features Profiling of Selected T Cell Epitopes

Some of the important features of the selected T-cell epitopes of both proteins were checked to support our study. Peptides are considered to be stable if digested by

fewer enzymes and are more favorable vaccine targets while the peptides that are digested by several enzymes are considered to be non-stable. Peptide digesting enzymes were predicted by Peptide cutter. Other features that include mutation, toxicity, hydrophobicity, PI, charge etc. were checked by multiple tools (Table 5.3 and 5.1).

4.3.5 Epitope Conservation Analysis

Multiple sequences of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN) and pre-protein translocase, SecG subunit from the same species were taken after performing the blast of these two proteins individually. Blast results into the 31 proteins for UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN) and 12 for the pre-protein translocase, SecG subunit that was subjected for multiple sequence alignment (Figure 4.16 and Figure 4.17). IEDB epitope conservancy analysis tool was used for the epitope conservation analysis of B-cell epitopes, MHC class-I and MHC class-II epitopes. The maximum identity shown by all the sequences is 100% for both proteins (Table 4.18 and 4.19).

TABLE 4.18: Conservation of epitopes of pre-protein translocase, SecG subunit via IEDB epitope conservancy analysis tool

Epitope sequence	**Epitope length**	**Percent of protein sequence matches at identity equal or greater 100%**	**Minimum identity**	**Maximum identity**
GFMGGGSE-TFYSKN	14	92.31% (12/13)	92.86%	100.00%
MHTFSIVLL	9	15.38% (2/13)	66.67%	100.00%
HTFSIVLLT	9	15.38% (2/13)	66.67%	100.00%
TFSIVLLTI	9	15.38% (2/13)	66.67%	100.00%
IVLLTIVSI	9	38.46% (5/13)	88.89%	100.00%
LLTIVSITL	9	92.31% (12/13)	88.89%	100.00%

IVLAQNLLS	9	38.46% (5/13)	66.67%	100.00%
IVSITLIVV	9	92.31% (12/13)	88.89%	100.00%
VVCSILFAI	9	53.85% (7/13)	88.89%	100.00%
ITLIVVVLM	9	92.31% (12/13)	88.89%	100.00%
LFAIIVLAQ	9	92.31% (12/13)	88.89%	100.00%
MQPSKTNGL	9	100.00% (13/13)	100.00%	100.00%
YSKNRTRTS	9	92.31% (12/13)	88.89%	100.00%
IVVVLMQPS	9	100.00% (13/13)	100.00%	100.00%
FAIIVLAQN	9	92.31% (12/13)	88.89%	100.00%
ILFAIIVLA	9	84.62% (11/13)	88.89%	100.00%
ITLIVVVLM	9	92.31% (12/13)	88.89%	100.00%

TABLE 4.19: Conservation of epitopes of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN) via IEDB epitope conservancy analysis tool

Epitope sequence	**Epitope length**	**Percent of protein sequence matches at identity equal or greater 100%**	**Minimum identity**	**Maximum identity**
TGTNGKT TTTSMIY	14	96.77% (30/31)	92.86%	100.00%
VENFSFN NKEEKYA	14	93.55% (29/31)	28.57%	100.00%
VKNPAGY DQAINTI	14	100.00% (31/31)	85.71%	100.00%
SNSTGAN MFPGIVA	14	87.10% (27/31)	57.14%	100.00%
EVDEANV KFITEYV	14	90.32% (28/31)	50.00%	100.00%
NDNYADG RDVSWIW	14	100.00% (31/31)	28.57%	100.00%
MCKINIKSF	9	100.00% (31/31)	33.33%	100.00%

CKINIKSFF	9	100.00% (31/31)	33.33%	100.00%
KINIKSFFS	9	100.00% (31/31)	33.33%	100.00%
FVISNSTGA	9	96.77% (30/31)	66.67%	100.00%
FSIIISKMV	9	74.19% (23/31)	33.33%	100.00%
IKSFFSIII	9	83.87% (26/31)	33.33%	100.00%
FPGIVACFV	9	87.10% (27/31)	55.56%	100.00%
LKLDKKILK	9	67.74% (21/31)	66.67%	100.00%
MFPGIVACF	9	90.32% (28/31)	55.56%	100.00%
YVSPEIITI	9	93.55% (29/31)	55.56%	100.00%
IEVDEANVK	9	100.00% (31/31)	66.67%	100.00%
FFSIIISKM	9	74.19% (23/31)	33.33%	100.00%
ILVTGTNGK	9	100.00% (31/31)	77.78%	100.00%
IISKMVLKL	9	74.19% (23/31)	33.33%	100.00%
FPGKVALKL	9	83.87% (26/31)	77.78%	100.00%
FNNKEEKYA	9	96.77% (30/31)	33.33%	100.00%
INIKSFFSI	9	100.00% (31/31)	33.33%	100.00%
YNMIKDSNK	9	100.00% (31/31)	55.56%	100.00%

FIGURE 4.16: Multiple sequence alignment of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN)

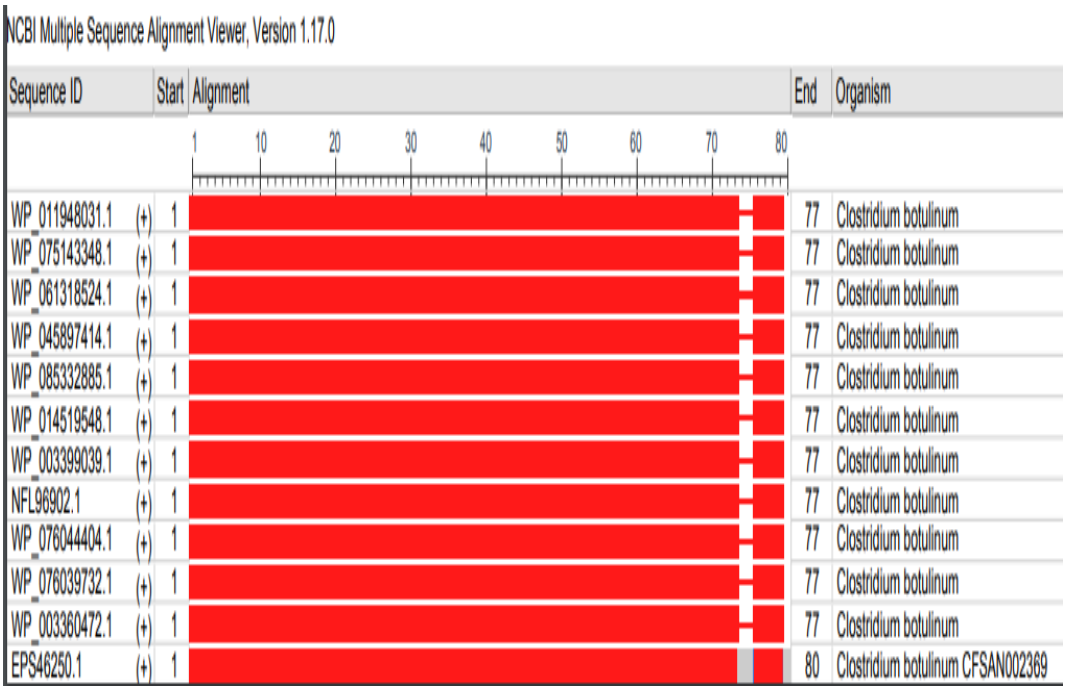


FIGURE 4.17: Multiple sequence alignment of pre-protein translocase, SecG subunit

4.3.6 Multi-epitope Vaccine Design and Construction

2 multi-epitope vaccines were designed based on the arrangements of peptides and GGGS linker was used for MHC Class-I and GPGPG linkers was used for joining MHC Class-II epitopes. The second vaccine contains B-defensin, an adjuvant and linker EAAAK at the N terminus. Figure 4.18 is showing the graphical representation of vaccines.

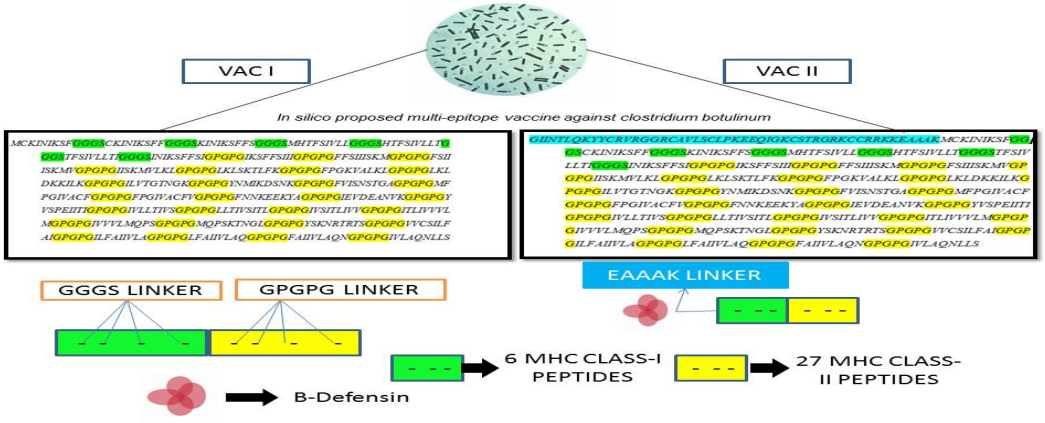


FIGURE 4.18: Graphical representation of VacI and VacII.

4.3.7 Physiochemical Properties of Multi-epitope Vaccines

Both the designed vaccines were projected for multiple physiochemical property analyses. AllerTop v 2.0 was used for allergenicity prediction and both the vaccines are considered as allergens. The antigenicity score of vaccine 1 is 1.0290 and score of vaccine 2 is 0.9995.

The molecular weight of vaccine 1 is 45412.93 KDa while of 2 is 51026.64, theoretical PI of vaccine 1 is 10.05 while of vaccine 2 is 10.07, instability index of vaccine 1 is 18.17 and of vaccine 2 is 21.81 that shows that both these vaccines are classified as stable vaccines.

Aliphatic index of vaccine 1 is 99.50 and of vaccine 2 is 96.28 while GRAVY of vaccine 1 is 0.516 and GRAVY of vaccine 2 is 0.401. vaccine 1 has M (Met)at N terminal meanwhile N terminal of Vaccine 2 has G (Gly).

But their estimated half-life is same that is half-life is 30 hours in mammalian reticulocytes if tested in vitro while if it is tested in-vivo it is greater than 20 hours in yeast and greater than 10 hours in Escherichia coli if tested in vivo.

4.3.8 Molecular Docking

Molecular docking was performed for 2 designed multi-epitope vaccines against 4 proteins with PDB ids as 3BO8 for HLA-A1, 6AT5.1.A for HLA-A2, 1BOR for HLA-A*0201 and 5IPF for HLA-A*0205.

Each protein is docked two times with 2 vaccines individually. 10 models for each docking job were downloaded and based on interactions only one model of vaccine with a protein was selected. Collectively eight models were selected which were presenting best interactions.

A PDBsum analysis tool is a tool that was used to analyze the residues involved in the protein-protein interaction. It gives the salt bridges, hydrogen bonds, disulphide bonds and non-bonded contacts.

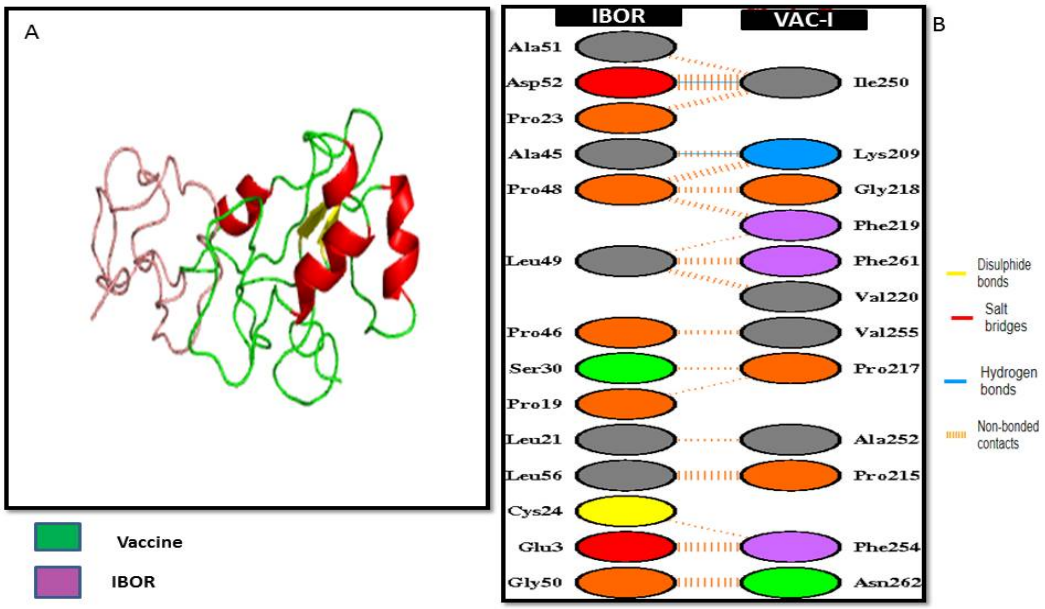


FIGURE 4.19: Protein-Protein interaction of 1BOR with Vac-1 (A) complex of Vac-I interacting with IBOR (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins.

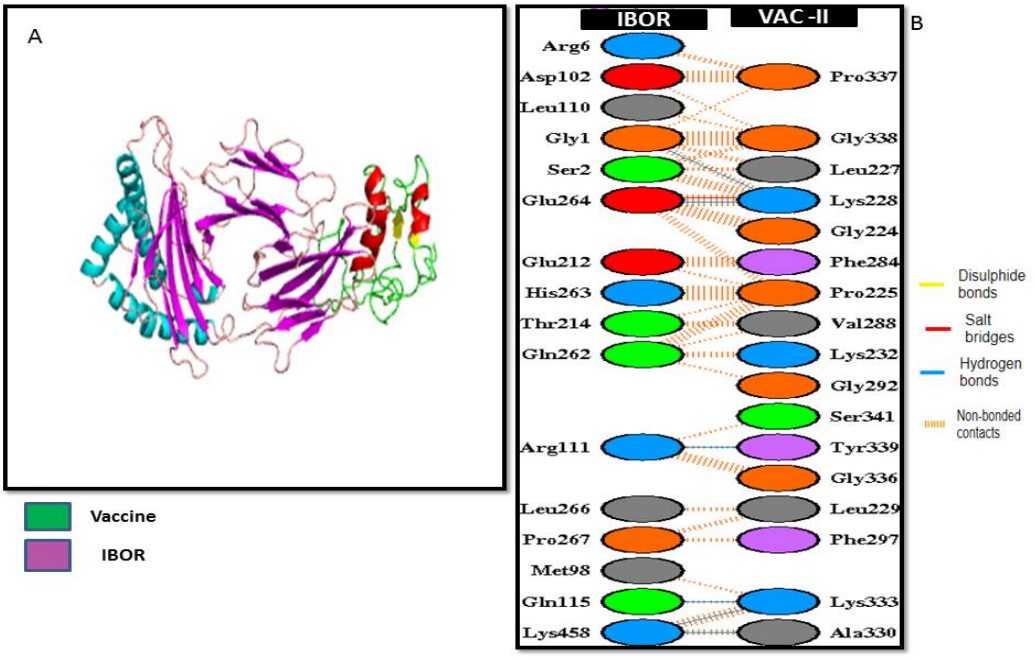


FIGURE 4.20: Protein-Protein interaction of 1BOR with Vac-II (A) complex of Vac-II interacting with IBOR (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins.

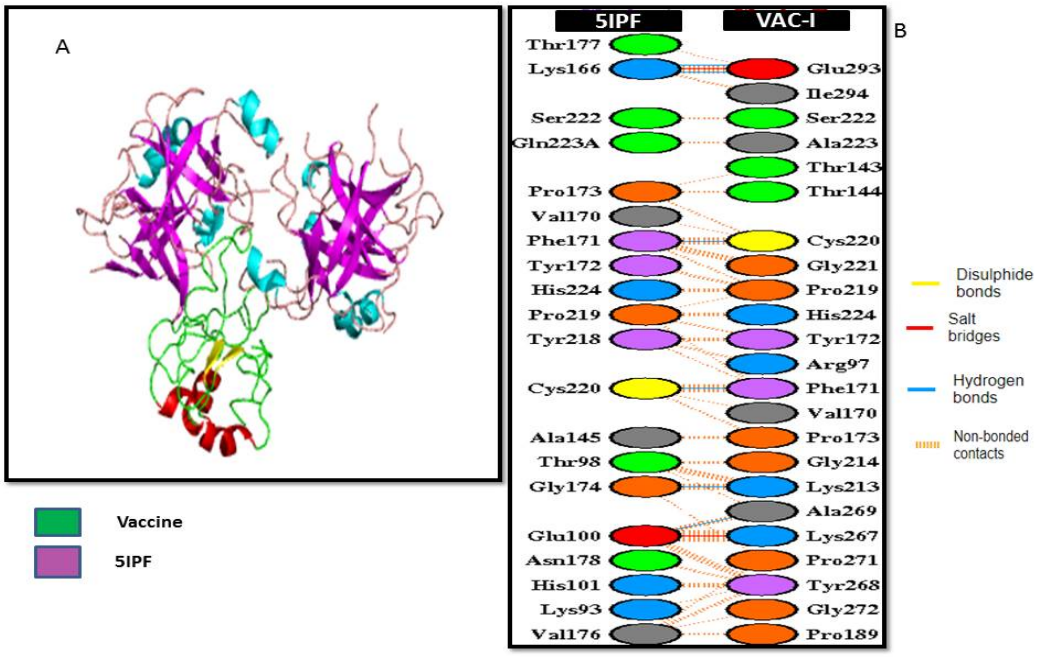


FIGURE 4.21: Protein-Protein interaction of 5IPF with Vac-I (A) complex of Vac-I interacting with 5IPF (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins.

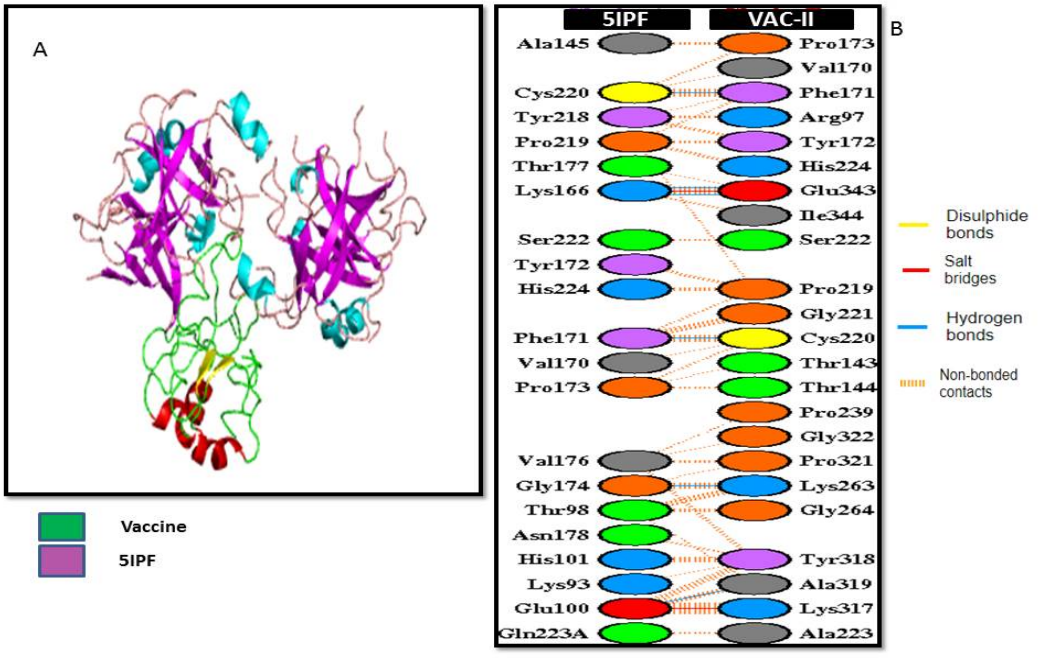


FIGURE 4.22: Protein-Protein interaction of 5IPF with Vac-II (A) complex of Vac-II interacting with 5IPF (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins.

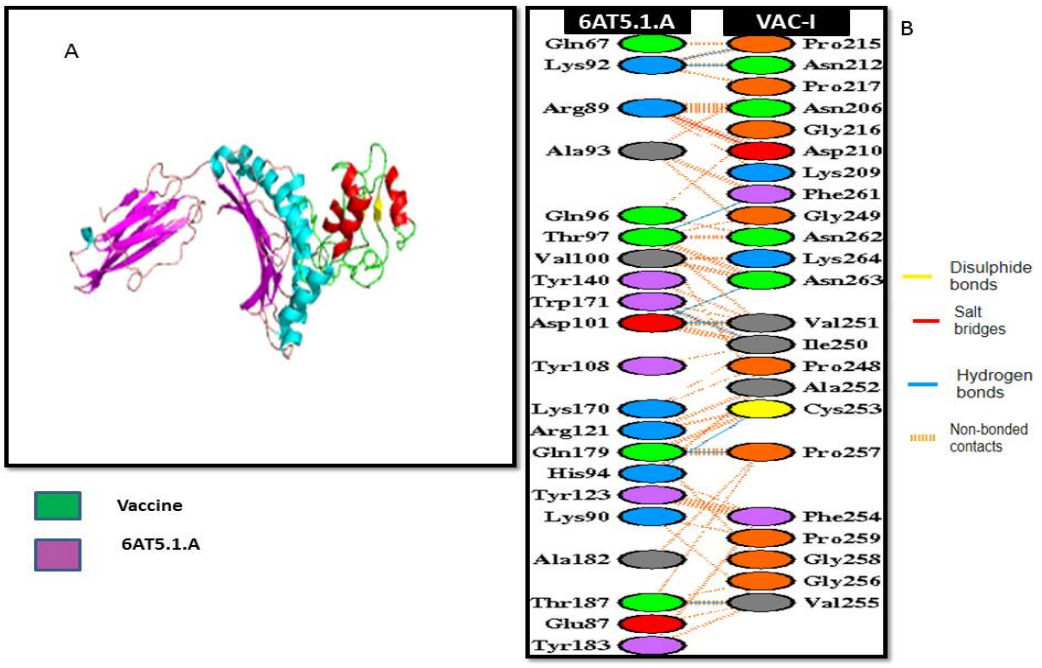


FIGURE 4.23: Protein-Protein interaction of 6AT5.1.A with Vac-I (A) complex of Vac-I interacting with 6AT5.1.A (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins.

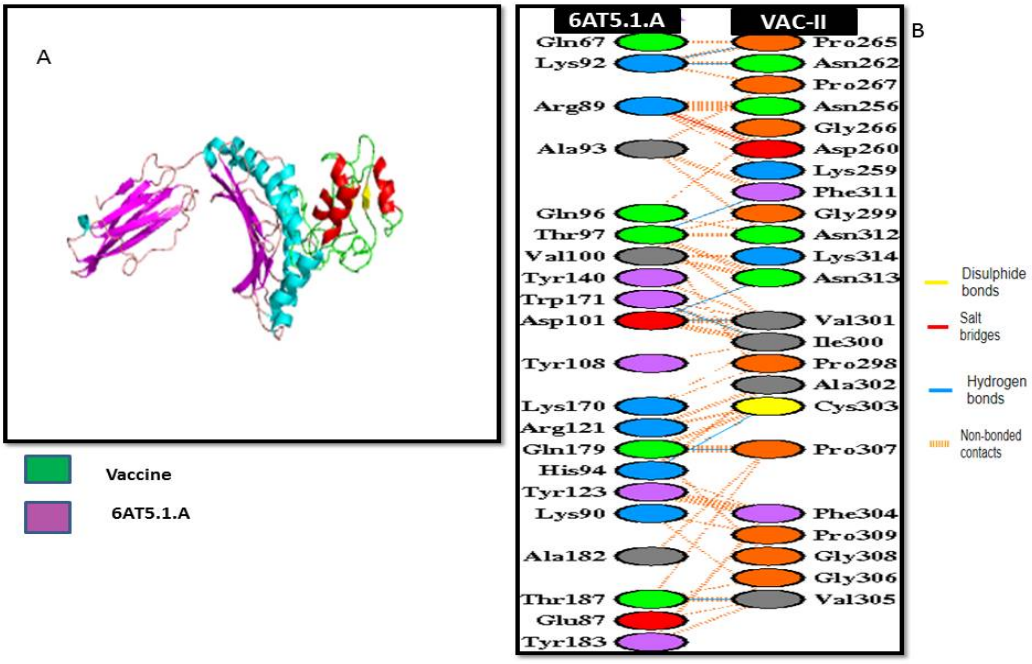


FIGURE 4.24: Protein-Protein interaction of 6AT5.1.A with Vac-II (A) complex of Vac-II interacting with 6AT5.1.A (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins.

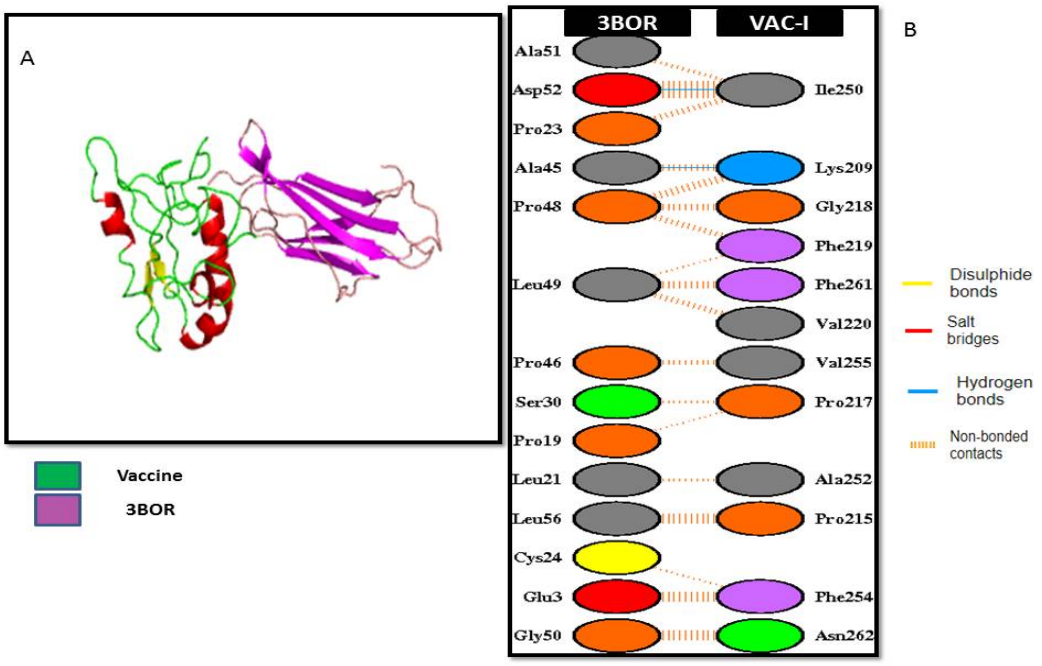


FIGURE 4.25: Protein-Protein interaction of 3BOR with Vac-I (A) complex of Vac-I interacting with 3BOR (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins.

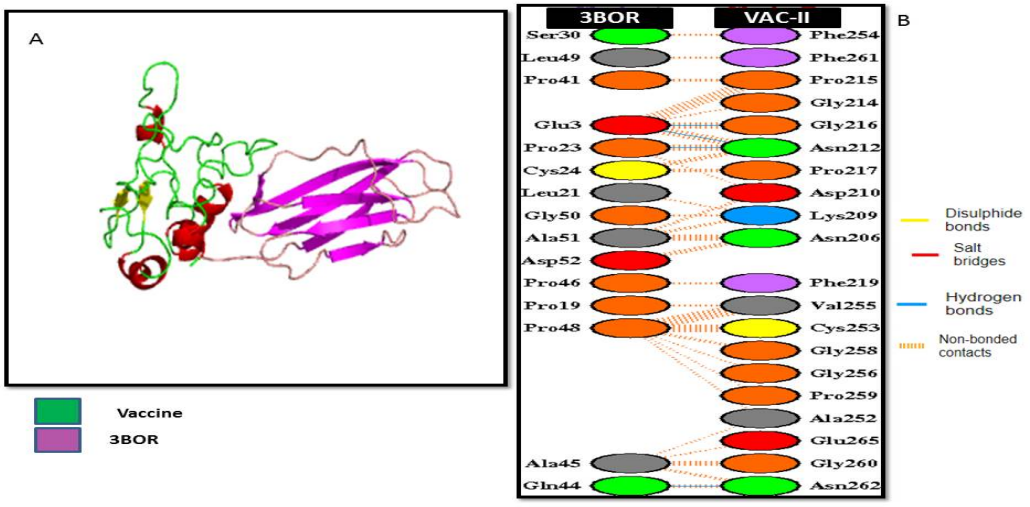


FIGURE 4.26: Protein-Protein interaction of 6AT5.1.A with Vac-II (A) complex of Vac-II interacting with 6AT5.1.A (B) Atomic-level interaction showing hydrogen bonds, non-bonded contacts, salt bridges and disulphide bonds among the residues of both interaction proteins.

Based on the interactions and PDBsum analysis, vaccine-II shows strong binding affinity and interactions with all the proteins.

Chapter 5

Conclusion and Future Prospects

C. botulinum is the cause of multiple deadly diseases worldwide that raises a major concern to find a solution that can stop this deadly bacteria to a great extent. The purpose of this study was to understand the genomic diversity of *C. botulinum* as well as to find novel drug and vaccine targets against this pathogen. 9 essential proteins were selected through the core-genome analysis that were used for this study. On the basis of cellular localization, these proteins were divided into membranous and cytoplasmic studies for drug targeting and epitope-based analysis. All the identified targets are playing an important role in the selected pathogen.

The first objective of this study was to explore the pan-genome and essential genome of *C. botulinum* and for this purpose 51 strains of *C. botulinum* were analyzed and core genome analysis was done for further study.

The second objective of this was to evaluate the potential of natural compounds against *C. botulinum* as a drug. For this purpose, 105 natural anti-bacterial compounds were collected through literature against 7 essential genes that were found through the cellular localization after drug prioritization. This is done in the drug-targeting approach. From this objective, 39 anti-bacterial compounds were shortlisted against all the proteins based on the binding affinity and residue interactions. The third objective was to identify the potential virulent factors in *C.*

botulinum and evaluation on the potential as drug and vaccine targets. For this purpose, 2 proteins were analyzed for the epitope-based study. 33 immunogenic B-cell derived T-cell epitopes of both the proteins pre-protein translocase, SecG subunit and UDP-N-acetylmuramyl tripeptide synthetase-like protein (Mur ligase family protein) because T-cells play an important role in the immune response of the body, were used for the formation of 2 vaccines, first with 2 types of linkers and second with the GGGS for MHC-Class I and GPGPG linkers was used for joining MHC Class-II epitopes. The second vaccine contains B-defensin, an adjuvant and linker EAAAK at the N terminus.

39 anti-bacterial compounds were selected based on their binding affinity as the lower the binding affinity, the better the interactions will be and the residue interaction with all the 7 essential cytoplasmic proteins. Both the vaccines docked against 4 different types of HLAs as they are an important part of the immune system. Top 8 models selected against 4 HLAs i.e., 3BO8 (HLA-A1), 6AT5.1. An (HLA-A2), 1BOR (HLA-Ax0201) and 5IPF (HLA-Ax0205) against 2 vaccines i.e., vaccine-I and vaccine-II based on interactions. Vaccine 2 shows better interaction with all the proteins that are checked by the PDBsum analysis of the docked models that gives the residue interaction of both proteins.

39 compounds for drug targets against *C. botulinum* that were selected can be further validated by in vitro analysis and can proceed for clinical trials. Apart from this, a reverse vaccinology approach was also used to find out the surface-exposed peptides that are the most efficient, cost-effective and less time-consuming approach. As there are no FDA approved vaccines are present against the botulinum toxins of 51 strains collectively. The epitope-based vaccine will enhance the immune response against all-natural infections. A lot of efforts are put into developing epitope-based vaccines because it provides the opportunity to engineer the combinations of epitopes as per requirements. This will also help us to facilitate the required immune responses on the T-cell epitopes. The data given in this study require further experimental authentication for verification but we anticipate promising outcomes from this predicted drug targets and peptide-epitopes against the deadly toxins of *C. botulinum*

Bibliography

[1] E. A. Johnson and M. Bradshaw, "Clostridium botulinum and its neurotoxins: a metabolic and cellular perspective," *Toxicon*, vol. 39, no. 11, pp. 1703–1722, 2001.

[2] M. W. Peck, "Biology and genomic analysis of clostridium botulinum," *Advances in microbial physiology*, vol. 55, pp. 183–320, 2009.

[3] M. Peck, "Clostridium botulinum and the safety of minimally heated, chilled foods: an emerging issue?" *Journal of Applied Microbiology*, vol. 101, no. 3, pp. 556–570, 2006.

[4] T. Bhardwaj and P. Somvanshi, "Pan-genome analysis of clostridium botulinum reveals unique targets for drug development," *Gene*, vol. 623, pp. 48–62, 2017.

[5] M. Naumann, L. M. Boo, A. H. Ackerman, and C. J. Gallagher, "Immunogenicity of botulinum toxins," *Journal of neural transmission*, vol. 120, no. 2, pp. 275–290, 2013.

[6] E. Dahlsten, H. Korkeala, P. Somervuo, and M. Lindström, "Pcr assay for differentiating between group i (proteolytic) and group ii (nonproteolytic) strains of clostridium botulinum," *International journal of food microbiology*, vol. 124, no. 1, pp. 108–111, 2008.

[7] G. Sakaguchi, "Clostridium botulinum toxins," *Pharmacology & therapeutics*, vol. 19, no. 2, pp. 165–194, 1982.

[8] D. Barh, K. Gupta, N. Jain, G. Khatri, N. León-Sicairos, A. Canizalez-Roman, S. Tiwari, A. Verma, S. Rahangdale, S. Shah Hassan *et al.*, "Conserved host–pathogen ppis globally conserved inter-species bacterial ppis based conserved host-pathogen interactome derived novel target in c. pseudotuberculosis, c. diphtheriae, m. tuberculosis, c. ulcerans, y. pestis, and e. coli targeted by piper betel compounds," *Integrative Biology*, vol. 5, no. 3, pp. 495–509, 2013.

[9] D. Perumal, C. S. Lim, K. R. Sakharkar, and M. K. Sakharkar, "Differential genome analyses of metabolic enzymes in pseudomonas aeruginosa for drug target identification," *In silico biology*, vol. 7, no. 4, 5, pp. 453–465, 2007.

[10] D. B. Thompson, K. Crandall, S. V. Harding, S. J. Smither, G. B. Kitto, R. W. Titball, and K. A. Brown, "In silico analysis of potential diagnostic targets from burkholderia pseudomallei," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 102, no. Supplement_1, pp. S61–S65, 2008.

[11] S. B. Jamal, S. S. Hassan, S. Tiwari, M. V. Viana, L. d. J. Benevides, A. Ullah, A. G. Turjanski, D. Barh, P. Ghosh, D. A. Costa *et al.*, "An integrative in-silico approach for therapeutic target identification in the human pathogen corynebacterium diphtheriae," *PloS one*, vol. 12, no. 10, p. e0186401, 2017.

[12] S. M. Asif, A. Asad, A. Faizan, M. S. Anjali, A. Arvind, K. Neelesh, K. Hirdesh, and K. Sanjay, "Dataset of potential targets for mycobacterium tuberculosis h37rv through comparative genome analysis," *Bioinformation*, vol. 4, no. 6, p. 245, 2009.

[13] Amir, S. M. Faizan, Asif, A. Asad, M. S. Anjali, A. Arvind, K. Neelesh, K. Hirdesh, and K. Sanjay, "Potential targets for mycobacterium tuberculosis h37rv through comparative genome analysis," *Bioinformation*, vol. 4, no. 6, p. 245, 2009.

[14] K. R. Sakharkar, M. K. Sakharkar, and V. T. Chow, "Biocomputational strategies for microbial drug target identification," in *New antibiotic targets*. Springer, 2008, pp. 1–9.

[15] B. Rathi, A. N. Sarangi, and N. Trivedi, "Genome subtraction for novel target definition in salmonella typhi," *Bioinformation*, vol. 4, no. 4, p. 143, 2009.

[16] D. Barh and A. Kumar, "In silico identification of candidate drug and vaccine targets from various pathways in neisseria gonorrhoeae," *In silico biology*, vol. 9, no. 4, pp. 225–231, 2009.

[17] J. N. Chan, C. Nislow, and A. Emili, "Recent advances and method development for drug target identification," *Trends in pharmacological sciences*, vol. 31, no. 2, pp. 82–88, 2010.

[18] D. Bumann, T. F. Meyer, and P. R. Jungblut, "Proteome analysis of the common human pathogen helicobacter pylori," *PROTEOMICS: International Edition*, vol. 1, no. 4, pp. 473–479, 2001.

[19] A. N. Sarangi, R. Aggarwal, Q. Rahman, and N. Trivedi, "Subtractive genomics approach for in silico identification and characterization of novel drug targets in neisseria meningitides serogroup b," *Journal of Computer Science & Systems Biology*, vol. 2, no. 5, pp. 255–258, 2009.

[20] M. J. Marton, J. L. DeRisi, H. A. Bennett, V. R. Iyer, M. R. Meyer, C. J. Roberts, R. Stoughton, J. Burchard, D. Slade, H. Dai *et al.*, "Drug target validation and identification of secondary drug target effects using dna microarrays," *Nature medicine*, vol. 4, no. 11, pp. 1293–1301, 1998.

[21] N. Chaturvedi and S. Nayak, "Future of generic drugs and india's interest," *Available at SSRN 1762356*, 2011.

[22] K. Mdluli and M. Spigelman, "Novel targets for tuberculosis drug discovery," *Current opinion in pharmacology*, vol. 6, no. 5, pp. 459–467, 2006.

[23] S. A. Plotkin, "Vaccines: the fourth century," *Clinical and Vaccine Immunology*, vol. 16, no. 12, pp. 1709–1719, 2009.

[24] M. Castelli, F. Cappelletti, R. A. Diotti, G. Sautto, E. Criscuolo, M. Dal Peraro, and N. Clementi, "Peptide-based vaccinology: experimental and computational approaches to target hypervariable viruses through the fine characterization of protective epitopes recognized by monoclonal antibodies and the identification of t-cell-activating peptides," *Clinical and Developmental Immunology*, vol. 2013, 2013.

[25] T. Ben-Yedidia and R. Arnon, "Epitope-based vaccine against influenza," *Expert review of vaccines*, vol. 6, no. 6, pp. 939–948, 2007.

[26] B. Pulendran and R. Ahmed, "Immunological mechanisms of vaccination," *Nature immunology*, vol. 12, no. 6, pp. 509–517, 2011.

[27] S. L. Demento, A. L. Siefert, A. Bandyopadhyay, F. A. Sharp, and T. M. Fahmy, "Pathogen-associated molecular patterns on biomaterials: a paradigm for engineering new vaccines," *Trends in biotechnology*, vol. 29, no. 6, pp. 294–306, 2011.

[28] C. D. Kane, J. E. Nuss, and S. Bavari, "Novel therapeutic uses and formulations of botulinum neurotoxins: A patent review (2012–2014)," *Expert opinion on therapeutic patents*, vol. 25, no. 6, pp. 675–690, 2015.

[29] R. K. Dhaked, M. K. Singh, P. Singh, and P. Gupta, "Botulinum toxin: bioweapon & magic drug," *The Indian journal of medical research*, vol. 132, no. 5, p. 489, 2010.

[30] J. Sobel, N. Tucker, A. Sulka, J. McLaughlin, and S. Maslanka, "Foodborne botulism in the united states, 1990–2000," *Emerging infectious diseases*, vol. 10, no. 9, p. 1606, 2004.

[31] T. Ono, T. Karashimada, and H. Iida, "Studies on the serum therapy of type e botulism (part iii)," *Japanese Journal of Medical Science and Biology*, vol. 23, no. 3, pp. 177–191, 1970.

[32] U. of Pennsylvania. Institute for Cooperative Research and H. E. Morton, *The Toxicity of Clostridium Botulinum Type: A Toxin for Various Species of Animals, Including Man*, 1961.

[33] C. for Disease Control, P. (US), C. for Disease Control, P. U. D. of Surveillance & Epidemiology. Systems Operation, and I. Branch, *Manual of Procedures for the Reporting of Nationally Notifiable Diseases to CDC.* US Department of Health and Human Services, Public Health Service, Centers. , 1995.

[34] G. Sakaguchi, "Clostridium botulinum toxins," *Pharmacology & Therapeutics*, vol. 19, no. 2, pp. 165–194, 1982. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0163725882900614

[35] L. S. Siegel, "Destruction of botulinum toxins in food and water," in *Clostridium botulinum.* CRC Press, 2018, pp. 323–341.

[36] L. M. Martinez-Levasseur, M. Simard, C. Furgal, G. Burness, P. Bertrand, S. Suppa, E. Avard, and M. Lemire, "Towards a better understanding of the benefits and risks of country food consumption using the case of walruses in nunavik (northern quebec, canada)," *Science of The Total Environment*, vol. 719, p. 137307, 2020.

[37] A. K. Rao, N. H. Lin, K. A. Jackson, R. K. Mody, and P. M. Griffin, "Clinical characteristics and ancillary test results among patients with botulism—united states, 2002–2015," *Clinical Infectious Diseases*, vol. 66, no. suppl_1, pp. S4–S10, 2018.

[38] Z. Beth, D. Hahn, B. Kramer, C. Tirre, D. Kruse, and P. A. Stone, "Open tibial fracture complicated by wound botulism: A case study," *The Journal of Foot and Ankle Surgery*, vol. 60, no. 3, pp. 600–604, 2021.

[39] A. Diaz and S. Sharma, "A case of a 34-year-old female with acute hypoxemic respiratory failure and proximal muscle weakness," *Case reports in critical care*, vol. 2017, 2017.

[40] A. H. Hauschild, "Epidemiology of human foodborne botulism," in *Clostridium botulinum*. CRC Press, 2018, pp. 69–104.

[41] R. Hinkle and N. Cox, "Infant botulism," *American Family Physician*, vol. 65, no. 7, p. 1388, 2002.

[42] R. Koepke, J. Sobel, and S. S. Arnon, "Global occurrence of infant botulism, 1976–2006," *Pediatrics*, vol. 122, no. 1, pp. e73–e82, 2008.

[43] T. F. Midura, "Update: infant botulism," *Clinical Microbiology Reviews*, vol. 9, no. 2, pp. 119–125, 1996.

[44] R. K. Dhaked, M. K. Singh, P. Singh, and P. Gupta, "Botulinum toxin: bioweapon & magic drug," *The Indian journal of medical research*, vol. 132, no. 5, p. 489, 2010.

[45] S. S. Arnon, R. Schechter, T. V. Inglesby, D. A. Henderson, J. G. Bartlett, M. S. Ascher, E. Eitzen, A. D. Fine, J. Hauer, M. Layton *et al.*, "Botulinum toxin as a biological weapon: medical and public health management," *Jama*, vol. 285, no. 8, pp. 1059–1070, 2001.

[46] T. Smith, C. H. Williamson, K. Hill, J. Sahl, and P. Keim, "Botulinum neurotoxin-producing bacteria. isn't it time that we called a species a species?" *MBio*, vol. 9, no. 5, pp. e01 469–18, 2018.

[47] M. Popoff, C. Mazuet, and B. Poulain, "Botulism and tetanus. the prokaryotes: human microbiology," 2013.

[48] H. Hariharan and W. Mitchell, "Observations on bacteriophages of clostridium botulinum type c isolates from different sources and the role of certain phages in toxigenicity," *Applied and environmental microbiology*, vol. 32, no. 1, pp. 145–158, 1976.

[49] W. Mitchell and Hariharan, "Observations on bacteriophages of clostridium botulinum type c isolates from different sources and the role of certain phages in toxigenicity," *Applied and environmental microbiology*, vol. 32, no. 1, pp. 145–158, 1976.

[50] K. Moriishi, M. Koura, N. Fujii, Y. Fujinaga, K. Inoue, B. Syuto, and K. Oguma, "Molecular cloning of the gene encoding the mosaic neurotoxin, composed of parts of botulinum neurotoxin types c1 and d, and pcr detection of this gene from clostridium botulinum type c organisms," *Applied and environmental microbiology*, vol. 62, no. 2, pp. 662–667, 1996.

[51] F. Defilippo, A. Luppi, G. Maioli, D. Marzi, M. C. Fontana, F. Paoli, P. Bonilauri, M. Dottori, and G. Merialdi, "Outbreak of type c botulism in birds and mammals in the emilia romagna region, northern italy," *Journal of wildlife diseases*, vol. 49, no. 4, pp. 1042–1046, 2013.

[52] A. Tanzi, O. McIntyre, A. Kolliopoulos, A. Rieu-Clarke, and R. Kinna, *The UNECE convention on the protection and use of transboundary watercourses and international lakes: its contribution to international water cooperation.* Hotei Publishing, 2015.

[53] A. Ventujol, A. Decors, C. l. Maréchal, J. Toux, V. Allain, C. Mazuet, M. Bayon-Auboyer, S. l. Bouquin, R. Souillard *et al.*, "Avian botulism in france: analysis of cases reported by two surveillance networks both in the wild and in poultry farms between 2000 and 2013." *Epidémiologie et Santé Animale*, no. 72, pp. 85–102, 2017.

[54] R. Souillard, C. Woudstra, C. Le Maréchal, M. Dia, M. Bayon-Auboyer, M. Chemaly, P. Fach, and S. Le Bouquin, "Investigation of clostridium botulinum in commercial poultry farms in france between 2011 and 2013," *Avian Pathology*, vol. 43, no. 5, pp. 458–464, 2014.

[55] P. Hambleton, "Clostridium botulinum toxins: a general review of involvement in disease, structure, mode of action and preparation for clinical use," *Journal of neurology*, vol. 239, no. 1, pp. 16–20, 1992.

[56] G. E. Hannett, W. B. Stone, S. W. Davis, and D. Wroblewski, "Biodiversity of clostridium botulinum type e associated with a large outbreak of botulism in wildlife from lake erie and lake ontario," *Applied and environmental microbiology*, vol. 77, no. 3, pp. 1061–1068, 2011.

[57] B. DasGupta and H. Sugiyama, "A common subunit structure in clostridium botulinum type a, b and e toxins," *Biochemical and biophysical research communications*, vol. 48, no. 1, pp. 108–112, 1972.

[58] S. Raffestin, B. Dupuy, J. C. Marvaud, and M. R. Popoff, "Botr/a and tetr are alternative rna polymerase sigma factors controlling the expression of the neurotoxin and associated protein genes in clostridium botulinum type a and clostridium tetani," *Molecular microbiology*, vol. 55, no. 1, pp. 235–249, 2005.

[59] J. O. Dolly, J. Black, R. S. Williams, and J. Melling, "Acceptors for botulinum neurotoxin reside on motor nerve terminals and mediate its internalization," *Nature*, vol. 307, no. 5950, pp. 457–460, 1984.

[60] S. Kozaki, "Interaction of botulinum type a, b and e derivative toxins with synaptosomes of rat brain," *Naunyn-Schmiedeberg's archives of pharmacology*, vol. 308, no. 1, pp. 67–70, 1979.

[61] P. Van den Abbeele, C. Grootaert, M. Marzorati, S. Possemiers, W. Verstraete, P. Gérard, S. Rabot, A. Bruneau, S. El Aidy, M. Derrien *et al.*, "Microbial community development in a dynamic gut model is reproducible, colon region specific, and selective for bacteroidetes and clostridium cluster ix," *Applied and environmental microbiology*, vol. 76, no. 15, pp. 5237–5246, 2010.

[62] J. Molgo and S. Thesleff, "Studies on the mode of action of botulinum toxin type a at the frog neuromuscular junction," *Brain research*, vol. 297, no. 2, pp. 309–316, 1984.

[63] B. Poulain, J. D. Wadsworth, E. A. Maisey, C. C. Shone, J. Melling, L. Tauc, and J. O. Dolly, "Inhibition of transmitter release by botulinum neurotoxin a: contribution of various fragments to the intoxication process," *European journal of biochemistry*, vol. 185, no. 1, pp. 197–203, 1989.

[64] R. Sriram, M. Shoff, G. Booton, P. Fuerst, and G. S. Visvesvara, "Survival of acanthamoeba cysts after desiccation for more than 20 years," *Journal of clinical microbiology*, vol. 46, no. 12, pp. 4045–4048, 2008.

[65] M. Adams, J. Kelley, J. Gocayne, M. P. Dubnick, and X. MH, "H. merril, cr wu, a. olde, b. moreno, rf kerlavage, ar mccombie, wr and venter. jc 1991," *Complementary DNA sequencing: Expressed sequence tags and human genome project. Science*, vol. 252, pp. 1651–1656.

[66] M. M. Gutacker, B. Mathema, H. Soini, E. Shashkina, B. N. Kreiswirth, E. A. Graviss, and J. M. Musser, "Single-nucleotide polymorphism–based population genetic analysis of mycobacterium tuberculosis strains from 4 geographic sites," *The Journal of infectious diseases*, vol. 193, no. 1, pp. 121–128, 2006.

[67] D. Barh, M. S. Khan, and E. Davies, *PlantOmics: the omics of plant science*. Springer, 2015.

[68] L. Zhang, D. Xiao, B. Pang, Q. Zhang, H. Zhou, L. Zhang, J. Zhang, and B. Kan, "The core proteome and pan proteome of salmonella paratyphi a epidemic strains," *PloS one*, vol. 9, no. 2, p. e89197, 2014.

[69] L. Rouli, V. Merhej, P. Fournier, and D. Raoult, "The bacterial pangenome as a new tool for analysing pathogenic bacteria. new microbes new infect 7: 72–85," 2015.

[70] J. W. Sahl, J. D. Gillece, J. M. Schupp, V. G. Waddell, E. M. Driebe, D. M. Engelthaler, and P. Keim, "Evolution of a pathogen: a comparative genomics analysis identifies a genetic pathway to pathogenesis in acinetobacter," *PloS one*, vol. 8, no. 1, p. e54287, 2013.

[71] I. Artin, A. T. Carter, E. Holst, M. Lovenklev, D. R. Mason, M. W. Peck, and P. Radstrom, "Effects of carbon dioxide on neurotoxin gene expression in nonproteolytic clostridium botulinum type e," *Applied and Environmental Microbiology*, vol. 74, no. 8, pp. 2391–2397, 2008.

[72] Y. Humeau, "Doussau f, grant nj, and poulain b," *How botulinum and tetanus neurotoxins block neurotransmitter release. Biochimie*, vol. 82, pp. 427–446, 2000.

[73] S. Sharma, M. Ramzan, and B. Singh, "Separation of the components of type a botulinum neurotoxin complex by electrophoresis," *Toxicon*, vol. 41, no. 3, pp. 321–331, 2003.

[74] S. A. Muhammad, S. Ahmed, A. Ali, H. Huang, X. Wu, X. F. Yang, A. Naz, and J. Chen, "Prioritizing drug targets in clostridium botulinum with a computational systems biology approach," *Genomics*, vol. 104, no. 1, pp. 24–35, 2014.

[75] M. R. Popoff, "Clostridial pore-forming toxins: powerful virulence factors," *Anaerobe*, vol. 30, pp. 220–238, 2014.

[76] H. Yoon, C. Ansong, J. E. McDermott, M. Gritsenko, R. D. Smith, F. Heffron, and J. N. Adkins, "Systems analysis of multiple regulator perturbations allows discovery of virulence factors in salmonella," *BMC systems biology*, vol. 5, no. 1, pp. 1–16, 2011.

[77] L. Zhao, L. Wong, and J. Li, "Antibody-specified b-cell epitope prediction in line with the principle of context-awareness," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 8, no. 6, pp. 1483–1494, 2011.

[78] L. Potocnakova, M. Bhide, and L. B. Pulzova, "An introduction to b-cell epitope mapping and in silico epitope prediction," *Journal of immunology research*, vol. 2016, 2016.

[79] J. Kubrycht, K. Sigler, and P. Souček, "Virtual interactomics of proteins from biochemical standpoint," *Molecular biology international*, vol. 2012, 2012.

[80] J. V. Kringelum, M. Nielsen, S. B. Padkjær, and O. Lund, "Structural analysis of b-cell epitopes in antibody: protein complexes," *Molecular immunology*, vol. 53, no. 1-2, pp. 24–34, 2013.

[81] E.-M. Yasser and V. Honavar, "Recent advances in b-cell epitope prediction methods," *Immunome research*, vol. 6, no. 2, pp. 1–9, 2010.

[82] J. D. Black and J. O. Dolly, "Interaction of 125i-labeled botulinum neurotoxins with nerve terminals. ii. autoradiographic evidence for its uptake into motor nerves by acceptor-mediated endocytosis." *The Journal of cell biology*, vol. 103, no. 2, pp. 535–544, 1986.

[83] M. Law, "Editorial overview: Preventive and therapeutic vaccines," *Current opinion in virology*, vol. 11, p. viii, 2015.

[84] Z. Chen, J. Li, and L. Wei, "A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue," *Artificial Intelligence in Medicine*, vol. 41, no. 2, pp. 161–175, 2007.

[85] T. R. Joaquim, R. Chambers, D. V. Onisk, F. Yin, J. M. Moriango, Y. Xu, D. A. Fancy, E. L. Crowgey, Y. He *et al.*, "Impact of immunization technology and assay application on antibody performance–a systematic comparative evaluation," *PloS one*, vol. 6, no. 12, p. e28718, 2011.

[86] M. C. Brown, T. R. Joaquim, R. Chambers, D. V. Onisk, F. Yin, J. M. Moriango, Y. Xu, D. A. Fancy, E. L. Crowgey, Y. He *et al.*, "Impact of immunization technology and assay application on antibody performance–a systematic comparative evaluation," *PloS one*, vol. 6, no. 12, p. e28718, 2011.

[87] N. A. Qureshi, S. M. Bakhtiar, M. Faheem, M. Shah, A. Bari, H. M. Mahmood, M. Sohaib, R. A. Mothana, R. Ullah, and S. B. Jamal, "Genome-based drug target identification in human pathogen streptococcus gallolyticus," *Frontiers in Genetics*, vol. 12, p. 303, 2021.

[88] A. M. Altenhoff, N. Škunca, N. Glover, C.-M. Train, A. Sueki, I. Piližota, K. Gori, B. Tomiczek, S. Müller, H. Redestig *et al.*, "The oma orthology database in 2015: function predictions, better plant support, synteny view and other improvements," *Nucleic acids research*, vol. 43, no. D1, pp. D240–D249, 2015.

[89] H. Luo, Y. Lin, F. Gao, C.-T. Zhang, and R. Zhang, "Deg 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements," *Nucleic acids research*, vol. 42, no. D1, pp. D574–D580, 2014.

[90] F. Agüero, B. Al-lazikani, M. Aslett, M. Berriman, S. Frederick, R. Campbell, S. Carmona, and I. Carruthers, "Edith a w.; chen f; et al," *Database*, vol. 7, no. 11, pp. 900–907, 2008.

[91] S. I. Mondal, S. Ferdous, N. A. Jewel, A. Akter, Z. Mahmud, M. M. Islam, T. Afrin, and N. Karim, "Identification of potential drug targets by subtractive genome analysis of escherichia coli o157: H7: an in silico approach," *Advances and applications in bioinformatics and chemistry: AABC*, vol. 8, p. 49, 2015.

[92] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "Kegg," *New perspectives*, vol. 855, 2008.

[93] M. S. Scott, S. J. Calafell, D. Y. Thomas, and M. T. Hallett, "Refining protein subcellular localization," *PLoS computational biology*, vol. 1, no. 6, p. e66, 2005.

[94] A. Volkamer, D. Kuhn, F. Rippmann, and M. Rarey, "Dogsitescorer: a web server for automatic binding site prediction, analysis and druggability assessment," *Bioinformatics*, vol. 28, no. 15, pp. 2074–2075, 2012.

[95] N. Tomar, V. Singh, S. Marla, R. Chandra, R. Kumar, and A. Kumar, "Molecular docking studies with rabies virus glycoprotein to design viral therapeutics," *Indian journal of pharmaceutical sciences*, vol. 72, no. 4, p. 486, 2010.

[96] C. Ibis, A. F. Tuyun, H. Bahar, S. S. Ayla, M. V. Stasevych, R. Y. Musyanovych, O. Komarovska-Porokhnyavets, and V. Novikov, "Synthesis of novel 1, 4-naphthoquinone derivatives: Antibacterial and antifungal agents," *Medicinal Chemistry Research*, vol. 22, no. 6, pp. 2879–2888, 2013.

[97] B. Shan, Y.-Z. Cai, J. D. Brooks, and H. Corke, "Antibacterial properties of polygonum cuspidatum roots and their major bioactive constituents," *Food Chemistry*, vol. 109, no. 3, pp. 530–537, 2008.

[98] E. Gasteiger, C. Hoogland, A. Gattiker, M. R. Wilkins, R. D. Appel, A. Bairoch *et al.*, "Protein identification and analysis tools on the expasy server," *The proteomics protocols handbook*, pp. 571–607, 2005.

[99] D. W. Buchan, F. Minneci, T. C. Nugent, K. Bryson, and D. T. Jones, "Scalable web services for the psipred protein analysis workbench," *Nucleic acids research*, vol. 41, no. W1, pp. W349–W357, 2013.

[100] F. Ferrè and P. Clote, "Dianna 1.1: an extension of the dianna web server for ternary cysteine classification," *Nucleic acids research*, vol. 34, no. suppl_2, pp. W182–W185, 2006.

[101] I. A. Doytchinova and D. R. Flower, "Vaxijen: a server for prediction of protective antigens, tumour antigens and subunit vaccines," *BMC bioinformatics*, vol. 8, no. 1, pp. 1–7, 2007.

[102] C. N. Magnan, A. Randall, and P. Baldi, "Solpro: accurate sequence-based prediction of protein solubility," *Bioinformatics*, vol. 25, no. 17, pp. 2200–2207, 2009.

[103] Y. EL-Manzalawy, D. Dobbs, and V. Honavar, "Predicting linear b-cell epitopes using string kernels," *Journal of Molecular Recognition: An Interdisciplinary Journal*, vol. 21, no. 4, pp. 243–255, 2008.

[104] J. M. Sánchez-Calvo, G. R. Barbero, G. Guerrero-Vásquez, A. G. Durán, M. Macías, M. A. Rodríguez-Iglesias, J. M. Molinillo, and F. A. Macías,

"Synthesis, antibacterial and antifungal activities of naphthoquinone derivatives: A structure–activity relationship study," *Medicinal Chemistry Research*, vol. 25, no. 6, pp. 1274–1285, 2016.

[105] M. C. Jespersen, B. Peters, M. Nielsen, and P. Marcatili, "Bepipred-2.0: improving sequence-based b-cell epitope prediction using conformational epitopes," *Nucleic acids research*, vol. 45, no. W1, pp. W24–W29, 2017.

[106] P. Koehl and M. Levitt, "Structure-based conformational preferences of amino acids," *Proceedings of the National Academy of Sciences*, vol. 96, no. 22, pp. 12 524–12 529, 1999.

[107] P. Sun, H. Ju, Z. Liu, Q. Ning, J. Zhang, X. Zhao, Y. Huang, Z. Ma, and Y. Li, "Bioinformatics resources and tools for conformational b-cell epitope prediction," *Computational and mathematical methods in medicine*, vol. 2013, 2013.

[108] X. Yang and X. Yu, "An introduction to epitope prediction methods and software," *Reviews in medical virology*, vol. 19, no. 2, pp. 77–96, 2009.

[109] N. London, B. Raveh, E. Cohen, G. Fathi, and O. Schueler-Furman, "Rosetta flexpepdock web server—high resolution modeling of peptide–protein interactions," *Nucleic acids research*, vol. 39, no. suppl_2, pp. W249–W253, 2011.

[110] E. A. Johnson and M. Bradshaw, "Clostridium botulinum and its neurotoxins: a metabolic and cellular perspective," *Toxicon*, vol. 39, no. 11, pp. 1703–1722, 2001.

[111] M. T. Ul Qamar, S. Saleem, U. A. Ashfaq, A. Bari, F. Anwar, and S. Alqahtani, "Epitope-based peptide vaccine design and target site depiction against middle east respiratory syndrome coronavirus: an immune-informatics study," *Journal of translational medicine*, vol. 17, no. 1, pp. 1–14, 2019.

[112] X. Chen, J. Zaro, and W.-C. Shen, "Fusion protein linkers: effects on production, bioactivity, and pharmacokinetics," *Fusion protein technologies for biopharmaceuticals: applications and challenges*, pp. 57–73, 2013.

[113] M. Saadi, A. Karkhah, and H. R. Nouri, "Development of a multi-epitope peptide vaccine inducing robust t cell responses against brucellosis using immunoinformatics based approaches," *Infection, Genetics and Evolution*, vol. 51, pp. 227–234, 2017.

[114] R. A. Shey, S. M. Ghogomu, K. K. Esoh, N. D. Nebangwa, C. M. Shintouo, N. F. Nongley, B. F. Asa, F. N. Ngale, L. Vanhamme, and J. Souopgui, "In-silico design of a multi-epitope vaccine candidate against onchocerciasis and related filarial diseases," *Scientific reports*, vol. 9, no. 1, pp. 1–18, 2019.

An Appendix

Table 5.1: Genomic Statistical features of selected strains of *Clostridium botulinum*

BioProject /GPID	**Assembly**	**Size (Mb)**	**GC %**	**Scaffolds**	**CDS**
PRJNA301982	GCA-001889325.1	4.39305	28	1	3992
PRJNA301977	GCA-001889365.1	4.43256	28.0622	2	4131
PRJNA302042	GCA-001921925.1	4.41435	28.0617	2	4111
PRJNA29859	GCA-000022765.1	4.15528	28.2	1	3792
PRJNA302048	GCA-002865845.1	4.15964	28.1979	2	3761
PRJNA301983	GCA-001886775.1	4.3522	27.9701	2	3895
PRJNA222585	GCA-000817935.1	4.36567	28.0293	2	3988
PRJNA301978	GCA-001879605.1	4.08903	28.1	1	3667
PRJNA302045	GCA-002865785.1	4.3189	27.9327	2	3878
PRJNA302041	GCA-002865885.1	4.08453	28.1793	3	3730

PRJNA302056	GCA-002866225.1	4.10915	28.1946	3	3769
PRJNA431675	GCA-003017195.1	4.00464	27.9	1	3564
PRJNA431675	GCA-003017225.1	3.99875	27.8	1	3561
PRJNA19519	GCA-000017065.1	4.01292	28.2904	2	3658
PRJNA302038	GCA-001921965.1	4.01184	28.1908	2	3571
PRJNA47575	GCA-000092345.1	4.01061	28.2904	2	3370
PRJNA28507	GCA-000019545.1	4.25969	28.1309	2	3794
PRJNA431675	GCA-003017145.1	3.98869	27.9	1	3558
PRJNA431675	GCA-003017335.1	3.97896	28.1	1	3543
PRJNA29077	GCA-000020345.1	4.25777	28.0255	3	3886
PRJNA302055	GCA-002866125.1	4.05048	28.1933	4	3670
PRJNA28505	GCA-000019305.1	4.10701	28.1949	2	3676
PRJNA302043	GCA-001921905.1	3.96772	28.1957	2	3551
PRJNA243331	GCA-003345315.1	4.0185	28.2062	3	3567
PRJNA301988	GCA-001879625.1	4.02006	28.205	3	3690

PRJNA302049	GCA-002865745.1	3.94513	28.2	1	3526
PRJNA302039	GCA-002866045.1	4.16951	28.1439	2	3829
PRJNA302044	GCA-001921945.1	4.19452	28.035	2	3754
PRJNA243331	GCA-003345335.1	4.20217	28.0232	3	3795
PRJNA302053	GCA-002865825.1	4.15495	28.0309	3	3799
PRJNA290487	GCA-003058445.1	3.93041	28.1954	2	3599
PRJEA61511	GCA-000253195.1	3.91974	28.2	1	3296
PRJNA302039	GCA-002865765.1	3.9176	28.3	1	3451
PRJNA302047	GCA-002865805.1	3.91457	28.1954	2	3457
PRJDB3402	GCA-000829015.1	3.9013	28.3	1	3577
PRJNA193	GCA-000063585.1	3.90326	28.1941	2	3523
PRJNA589706	GCA-009733885.1	3.90311	28.1941	2	3670
PRJNA225787	GCA-000789355.1	3.9146	27.4103	2	3478
PRJNA301979	GCA 001889345.1	3.86763	28.2	1	3362
PRJNA19517	GCA-000017025.1	3.86345	28.2	1	3627

PRJNA290487	GCA-003058345.1	3.85851	28.2	1	3587
PRJNA302036	GCA-001921985.1	4.30335	27.8925	4	3970
PRJNA28857	GCA-000020165.1	3.84797	27.469	2	3427
PRJEA60761	GCA-000307125.1	3.8292	27.3701	2	3360
PRJNA19521	GCA-000017045.1	3.76056	28.2	1	3327
PRJNA28855	GCA-000020285.1	3.65964	27.4	1	3130
PRJNA272602	GCA-000827955.1	3.6119	27.4	1	3068
PRJNA272603	GCA-000827935.1	3.6119	27.4	1	3071
PRJNA60407	GCA-000204565.1	3.20759	28.2433	6	2919

TABLE 5.2: MHC class-II allele binding peptides of pre-protein translocase, SecG subunit predicted via Propred with their antigenicity scores

Peptide sequence	**Position**	**MHC class II alleles**	**Vaxijen score**

IVLLTIVSI	5	DRB1-0101, DRB1-0102, DRB1-0301, DRB1-0305, DRB1-0306, DRB1-0307, DRB1-0308, DRB1-0309, DRB1-0311, DRB1-0401, DRB1-0404, DRB1-0405, DRB1-0408, DRB1-0410, DRB1-0421, DRB1-0423, DRB1-0426, DRB1-0701, DRB1-0703, DRB1-0801, DRB1-0802, DRB1-0804, DRB1-0806, DRB1-0813, DRB1-0817, DRB1-1101, DRB1-1102, DRB1-1104, DRB1-1106, DRB1-1107, DRB1-1114, DRB1-1120, DRB1-1121, DRB1-1128, DRB1-1301, DRB1-1302, DRB1-1304, DRB1-1305, DRB1-1307, DRB1-1311, DRB1-1321, DRB1-1322, DRB1-1323, DRB1-1327, DRB1-1328, DRB1-1502, DRB1-1506	0.4421
LLTIVSITL	7	DRB1-0101, DRB1-0102, DRB1-0701, DRB1-0703, DRB1-1501, DRB1-1502, DRB1-1506, DRB5-0105	0.5297

IVLAQNLLS	67	DRB1-0101, DRB1-0102, DRB1-0301, DRB1-0305, DRB1-0306, DRB1-0307, DRB1-0308, DRB1-0309, DRB1-0311, DRB1-0401, DRB1-0402, DRB1-0404, DRB1-0405, DRB1-0408, DRB1-0410, DRB1-0421, DRB1-0423, DRB1-0426, DRB1-0802, DRB1-0804, DRB1-0806, DRB1-0813, DRB1-0817, DRB1-1101, DRB1-1102, DRB1-1104, DRB1-1106, DRB1-1107, DRB1-1114, DRB1-1120, DRB1-1121, DRB1-1128, DRB1-1301, DRB1-1302, DRB1-1304, DRB1-1305, DRB1-1307, DRB1-1311, DRB1-1321, DRB1-1322, DRB1-1323, DRB1-1327, DRB1-1328, DRB1-1501, DRB1-1502, DRB1-1506	0.1512
IVSITLIVV	10	DRB1-0101, DRB1-0102, DRB1-0301, DRB1-0309, DRB1-0701, DRB1-0703, DRB1-1101, DRB1-1104, DRB1-1106, DRB1-1128, DRB1-1305, DRB1-1311, DRB1-1321, DRB1-1501, DRB1-1502, DRB1-1506, DRB5-0105	0.5512
VVCSILFAI	58	DRB1-0301, DRB1-0305, DRB1-0309, DRB1-0701, DRB1-0703, DRB1-1107, DRB1-1301, DRB1-1302, DRB1-1327, DRB1-1328, DRB1-1501, DRB1-1502, DRB1-1506	1.5096
ITLIVVVLM	13	DRB1-0301, DRB1-0309, DRB1-0405, DRB1-0410, DRB1-0421, DRB1-1305, DRB1-1327, DRB1-1328	0.7628

LFAIIVLAQ	63	DRB1-0305, DRB1-0306, DRB1-0307, DRB1-0308, DRB1-0311, DRB1-0401, DRB1-0402, DRB1-0404, DRB1-0405, DRB1-0408, DRB1-0410, DRB1-0423, DRB1-0426, DRB1-0801, DRB1-0802, DRB1-0804, DRB1-0806, DRB1-0817, DRB1-1101, DRB1-1102, DRB1-1104, DRB1-1106, DRB1-1107, DRB1-1121, DRB1-1128, DRB1-1304, DRB1-1305, DRB1-1307, DRB1-1311, DRB1-1321, DRB1-1322	0.5012
MQPSKTNGL	21	DRB1-0306, DRB1-0307, DRB1-0308, DRB1-0311, DRB1-0401, DRB1-0421, DRB1-0426, DRB1-0701, DRB1-0703	0.5515
YSKNRTRTS	41	DRB1-0401, DRB1-0402, DRB1-0421, DRB1-0426, DRB1-1114, DRB1-1302, DRB1-1323	0.7072
IVVVLMQPS	16	DRB1-0402, DRB5-0105	0.3367
FAIIVLAQN	64	DRB1-0801, DRB1-1307, DRB1-1321	0.2089
ILFAIIVLA	62	DRB1-1102, DRB1-1114, DRB1-1121, DRB1-1322, DRB1-1323	1.4669
ITLIVVVLM	13	DRB1-1128, DRB1-1301	0.7628

TABLE 5.3: MHC class-II allele binding peptides of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN) predicted via Propred with their antigenicity scores

Peptide sequence	**Position**	**MHC class II alleles**	**Vaxijen score**

FVISNSTGA	79	DRB1-0101, DRB1-0102, DRB1-0305, DRB1-0306, DRB1-0307, DRB1-0308, DRB1-0309, DRB1-0311, DRB1-0401, DRB1-0402, DRB1-0404, DRB1-0405, DRB1-0408, DRB1-0410, DRB1-0421, DRB1-0423, DRB1-0426, DRB1-0701, DRB1-0703, DRB1-0813, DRB1-1101, DRB1-1107, DRB1-1114, DRB1-1120, DRB1-1302, DRB1-1307, DRB1-1323, DRB5-0101, DRB5-0105	0.8079
FSIIISKMV	9	DRB1-0101, DRB1-0102, DRB1-0401, DRB1-0426, DRB1-0701, DRB1-0703, DRB1-1307	0.7007
IKSFFSIII	5	DRB1-0101, DRB1-0102, DRB1-0402, DRB1-0404, DRB1-0405, DRB1-0410, DRB1-0423, DRB1-0701, DRB1-0703, DRB1-1104, DRB1-1106, DRB1-1311, DRB1-1501, DRB1-1502, DRB1-1506	0.6648
FPGIVACFV	90	DRB1-0101	0.4499
LKLDKKILK	37	DRB1-0301, DRB1-0305, DRB1-0306, DRB1-0307, DRB1-0308, DRB1-0309, DRB1-0311, DRB1-1104, DRB1-1106, DRB1-1107, DRB1-1311	0.7045
MFPGIVACF	89	DRB1-0301	0.5690

LKLSKTLFK	18	DRB1-0301, DRB1-0305, DRB1-0306, DRB1-0307, DRB1-0308, DRB1-0311, DRB1-0401, DRB1-0402, DRB1-0404, DRB1-0408, DRB1-0410, DRB1-0421, DRB1-0423, DRB1-0426, DRB1-0804, DRB1-1101, DRB1-1102, DRB1-1104, DRB1-1106, DRB1-1107, DRB1-1114, DRB1-1121, DRB1-1128, DRB1-1305, DRB1-1307, DRB1-1311, DRB1-1322, DRB1-1323, DRB5-0101, DRB5-0105	0.2046
YVSPEIITI	126	DRB1-0305, DRB1-0309, DRB1-0701, DRB1-0703, DRB1-0813, DRB1-0817, DRB1-1101, DRB1-1128, DRB1-1305, DRB1-1321, DRB1-1501, DRB1-1502, DRB1-1502, DRB1-1506	0.5986
IEVDEANVK	113	DRB1-0306, DRB1-0307, DRB1-0308, DRB1-0311	2.0926
FFSIIISKM	8	DRB1-0309, DRB1-0405, DRB1-0408, DRB1-0801, DRB1-0813, DRB1-1128, DRB1-1305, DRB1-1307, DRB1-1321	0.8452
ILVTGTNGK	54	DRB1-0401, DRB1-0402, DRB1-0404, DRB1-0421, DRB1-0423, DRB1-0426, DRB5-0101, DRB5-0105	1.1673
IISKMVLKL	12	DRB1-0801, DRB1-0802, DRB1-0804, DRB1-0806, DRB1-0813, DRB1-0813, DRB1-0817, DRB1-1101, DRB1-1102, DRB1-1104, DRB1-1106, DRB1-1114, DRB1-1120, DRB1-1121, DRB1-1128, DRB1-1302, DRB1-1304, DRB1-1305, DRB1-1311, DRB1-1321, DRB1-1322, DRB1-1323, DRB1-1327, DRB1-1328	0.7117

FPGKVALKL	31	DRB1-0801, DRB1-0813, DRB1-0817	0.9355
FNNKEEKYA	103	DRB1-0802	0.7212
INIKSFFSI	3	DRB1-0802, DRB1-0804, DRB1-0806, DRB1-1102, DRB1-1121, DRB1-1301, DRB1-1304, DRB1-1322, DRB1-1327, DRB1-1328, DRB1-1501, DRB1-1502, DRB1-1506	0.9882
YNMIKDSNK	70	DRB5-0101, DRB5-0105	0.0666

Table 5.4: Biological and molecular characteristics of the essential proteins

Uniprot Id	**Molecular Function**	**Biological Function**	**Protein name**	**Gene Name**	**Cellular Localization**	**Virulencec**	**Pathways**
A5HY47 (ATPL_CLOBH)	lipid binding	ATP synthesis coupled proton transport	ATP synthase subunit c	atpE	Innermembrane, cytoplasmic, periplasmic	No	adenosine ribonucleotides de novo biosynthesis
	proton-transporting ATP synthase activity, rotational mechanism						
A5HXP7	ATP binding	DNA replication initiation	Chromosomal replication initiator protein DnaA	dnaA	Cytoplasmic	Yes	Two-component system

	DNA replication origin binding	regulation of DNA replication					
A5HY25 (A5HY25_CLOBH)	ATP binding	thiamine biosynthetic process	ThiI domain-containing protein	CBO0-129	Cytoplasmic	Yes	Sulfur relay system
	sulfurtransferase activity	thiamine diphosphate biosynthetic process					
	tRNAadenylyl-transferase activity	thiazole biosynthetic process					
	tRNA binding	tRNA 4-thiouridine biosynthesis					

A5HY78 (A5HY78_CLOBH)	GTP diphosphokinase activity	guanosine tetraphosphate metabolic process	Putative guanosine 3',5'-bispyrophosphate (PpGpp) synthesis/degradation protein	CBO-0185	Cytoplasmic	Yes	Purine metabolism, metabolic pathways

A5I0U0 (A5I0U0_CLOBH)	Hydrolase activity	RNA repair tRNA 3'-terminal CCA addition	Putative metal dependent phosphohydrolase	CBO-1098	Cytoplasmic	Yes	Metabolic pathways, Biosynthesis of secondary metabolites, Biosynthesis of antibiotics, Fatty acid biosynthesis, Carbon metabolism, Fatty acid metabolism, Pyruvate metabolism, Propanoate metabolism, Aminoacyl-tRNA biosynthesis, Microbial metabolism in diverse environments

A5I1T7 (A5I1T7_CLOBH)	ATP binding	cell cycle	UDP-N-acetylmuramoyltripeptide–D-alanyl-D-alanine ligase	murF	Cytoplasmic	Yes	Vancomycin resistance, Peptidoglycan biosynthesis, Metabolic pathways, Lysine biosynthesis
	UDP-N-acetylmuramoylalanyl-D-glutamate-2,6-diaminopimelate ligase activity	cell division					
		cell wall organization peptidoglycan biosynthetic process regulation of cell shape					

A5I2P9 (A5I2P9_CLOBH)	DNA binding	phosphorelay signal transduction system regulation of transcription, DNA-templated	Chemotaxis protein	cheV	Cytoplasmic	Yes	Two-component system, Bacterial chemotaxis

*Molecular function (MF) and biological process (BP) for each target protein were determined using UniProt.

*Molecular function (MF) and biological process (BP) for each target protein were determined using UniProt.

**Cellular localization of pathogen targets was performed using CELLO.

***Virulence was find out by PAIDB and VFDB.

****Molecular weight was determined using ProtParam tool (http://web.expasy. org/protparam/).

*****KEGG was used to find the role of these targets in different cellular pathways.

TABLE 5.5: Peptides of UDP-N-acetylmuramyl tripeptide synthetase-like protein (MUR LIGASE FAMILY PROTEIN) with non-digesting enzymes, mutation position, toxicity, allergenicity, hydrophobicity, hydrophilicity, charge, and PI.

Peptide	**Non-digesting Enzymes**	**Mutation position**	**Toxicity**	**Allergenicity**	**Hydrophobicity**	**Hydrophilicity**	**Charge**	**PI**
			MHC Class 1 Alleles					
MCKINIKSF	Arg-C proteinase, Asp-Nendopeptidase, Caspase1-9, Enterokinase, Glutamyl endopeptidase, GranzymeB, Proline-endopeptidase, Staphylococcal peptidase I, Thrombin Tobacco etch	NM	NT	A	-0.08	-0.21	2.00	9.36

CKINIKSFF	Arg-C proteinase, Asp-N endopeptidase,Caspase1 -9, Enterokinase, Glutamyl endopeptidase, GranzymeB, Proline-endopeptidase, Staphylococcal peptidase I, Thrombin, Tobacco etc h virus protease	NM	NT	A	-0.04	-0.34	2.00	9.36

KINIKSFFS	Arg-C proteinase, Asp-N endopeptidase,Caspase1-9,Enterokinase, Glutamyl endopeptidase, GranzymeB, Proline-endopeptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease	NM	NT	N/A	-0.08	-0.20	2.00	10.02

MHC Class 2 Alleles

FVISNSTGA	Arg-C proteinase, Asp-N endo-peptidase, Caspase1-9, Enterokinase, Glutamyl endo-peptidase, GranzymeB, Proline-endo-peptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease	NM	NT	N/A	0.11	-0.66	0.00	5.88

FSIIISKMV	Arg-C proteinase, Asp-N endo-peptidase, Caspase1-9, Enterokinase, Glutamyl endo-peptidase, GranzymeB, Proline-endo-peptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease, Clostripain	NM	NT	N/A	0.22	-0.79	1.00	9.11

IKSFFSIII	Caspase1-9, Enterokinase, Glutamyl endo-peptidase, GranzymeB, Proline-endo-peptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease, Clostripain, Arg-C protein-ase,Asp-N endo-peptidase, Neutrophil elastase	NM	NT	N/A	0.28	-0.96	1.00	9.11

FPGIVACFV	Arg-C proteinase, Asp-N endo-peptidase, Caspase1-9, Clostripain, Enterokinase, Glutamyl endo-peptidase, GranzymeB, LysC, LysN, Proline-endo-peptidase, Staphylococcal peptidase I, Thrombin, Trypsin, Tobacco etch virus protease	NM	NT	N/A	0.38	-1.26	0.00	5.85

LKLDKKILK	Arg-C proteinase, Caspase1-10, Chymotrypsin-high specificity, Clostripain, Enterokinase, Glutamyl endo-peptidase, GranzymeB, Neutrophil elas-tase, Proline-endopeptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease	NM	NT	A	-0.31	0.87	3.00	10.01

MFPGIVACF	Arg-C proteinase, Asp-N endo-peptidase, Caspase1-10, Enterokinase, Glutamyl endo peptidase, GranzymeB, Neutrophil elastase, Proline -endopeptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease, Trypsin	NM	NT	A	0.35	-1.23	0.00	5.85

YVSPEIITI	Arg-C proteinase, Asp-N endo-peptidase, Caspase1-10, Enterokinase, Glutamyl endo-peptidase, GranzymeB, Proline-endo-peptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease, Trypsin, LySC, LySN	NM	NT	A	0.18	-0.70	-1.00	4.00

IEVDEANVK	Arg-C proteinase, Caspase1-10, Chymotrypsin-high specificity, Chymotrypsin-low specificity, Clostripain, Enterokinase, GranzymeB Pepsin (pH1.3), Pepsin (pH2), Proline-endo-peptidase, Thrombin, Tobacco etch virus protease	NM	NT	A	-0.18	0.77	-2.00	4.14

FFSIIISKM	Arg-C proteinase, Asp-N endo-peptidase, Caspase 1-10, Enterokinase, Glutamyl endo-peptidase, GranzymeB, Neutrophil elastase, Proline-endo-peptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease	NM	NT	A	0.23	-0.90	1.00	9.11

ILVTGTNGK	Arg-C proteinase, Asp-N endo-peptidase, Caspase1-10, Chymotrypsin-high specificity, Clostripain, Enterokinase, Glutamyl endo-peptidase, GranzymeB, Proline-endo-peptidase, Staphylococcal peptidase I, Thrombin tobacco etch virus protease	NM	NT	N/A	0.00	-0.30	1.00	9.11

IISKMVLKL	Arg-C proteinase, Asp-N endo-peptidase, Caspase1-10, Chymotrypsin-high specificity, Clostripain, Enterokinase, Glutamyl endopeptidase, GranzymeB, Proline-endo-peptidase, Staphylococcal peptidase I, Thrombin tobacco etch virus protease	NM	NT	A	0.10	-0.41	2.00	10.02

FPGKVALKL	Arg-C proteinase, Asp-N endo-peptidase, Caspase1-10, Chymotrypsin-high specificity, Clostripain, Enterokinase, Glutamyl endo-peptidase, GranzymeB, Proline-endo-peptidase, Staphylococcal peptidase I, Thrombin tobacco etch virus protease	NM	NT	N/A	0.04	-0.23	2.00	10.02

FNNKEEKYA	Arg-C proteinase, Asp-N endo-peptidase, Caspase1-10, Clostripain, Enterokinase, GranzymeB, Proline-endo-peptidase, Thrombin, Tobacco etch virus protease	NM	NT	A	-0.43	0.79	0.00	6.49

INIKSFFSI	Arg-C proteinase, Asp-N endo-peptidase, Caspase1-10, Chymotrypsin-high specificity, Clostripain, Enterokinase, Glutamyl endo-peptidase, GranzymeB, Proline-endo-peptidase, Staphylococcal peptidase I, Thrombin tobacco etch virus protease	NM	NT	N/A	0.13	-0.73	1.00	9.11

YNMIKDSNK	Arg-C proteinase, Asp-N endo-peptidase, Caspase1-10, Chymotrypsin-high specificity, Clostripain, Enterokinase, Glutamyl endo-peptidase,	NM	NT	A	-0.38	0.48	1.00	8.33

NA: NON-ALLERGEN

A: ALLERGEN

NM: NO MUTATION

NT: NON-TOXIN

TABLE 5.6: Peptides of pre-protein translocase, SecG subunit with non-digesting enzymes, mutation position, toxicity, allergenicity, hydrophobicity, hydrophilicity, charge, and PI

Peptide	**Non-digesting Enzymes**	**Mutation position**	**Toxicity**	**Allergenicity**	**Hydrophobicity**	**Hydrophilicity**	**Charge**	**PI**
			MHC Class 1 Alleles					
MHTFSIVLL	Arg-C proteinase, Asp-N endo-Clostripain, Enterokinase Glutamyl endo-peptidase, GranzymeB, peptidase I, Thrombin, Tobacco etch virus protease, Trypsin	NM	NT	N/A	0.26	-1.26	0.60	7.10

HTFSIVLLT	Arg-C proteinase, Asp-N endo-peptidase, caspases 1-10, Clostripain, Enterokinase, Glutamyl endo-peptidase, GranzymeB, LysC, LysN, Proline-endo-peptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease, Trypsin	NM	NT	A	0.21	-1.16	0.50	7.10

TFSIVLLTI	Arg-C proteinase, Asp-N endo-peptidase, caspases 1-10, Clostripain, Enterokinase Glutamyl endo-peptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease, Trypsin	NM	NT	A	0.34	-1.30	0.00	5.88

MHC Class 2 Alleles

IVLLTIVSI	Arg-C proteinase, Asp-N endo-peptidase, caspases 1-10, Chymotrypsin high specificity, Clostripain, Enterokinase, Glutamyl endopeptidase, GranzymeB, LysC, LysN, Proline-endo-peptidase, virus protease, Trypsin	NM	NT	A	0.43	-1.34	0.00	5.88

LLTIVSITL	Arg-C proteinase, Asp-N endo-peptidase, caspases 1-10, Chymotrypsin-high specificity, Clostripain, Enterokinase, Glutamyl endo-peptidase, virus protease, Trypsin	NM	NT	A	0.33	-1.22	0.00	5.88

IVLAQNLLS	Arg-C proteinase, Asp-N endopeptidase Asp-N endopeptidase + N-terminal Glu, caspases 1-10, Chymotrypsin-high specificity, Clostripain, Enterokinase, Tobacco etch virus protease, Trypsin	NM	NT	A	0.17	-0.94	0.00	5.88

IVSITLIVV	Arg-C proteinase, Asp-N endo-peptidase, Asp-N endopeptidase + N-terminal Glu, caspases 1-10, Chymotrypsin-high specificity, Clostripain, Enterokinase, Glutamyl endo-peptidase, Thrombin, Tobacco etch virus protease, Trypsin	NM	NT	A	0.43	-1.31	0.00	5.88

VVCSILFAI	Arg-C proteinase, Asp-N endo-peptidase Asp-N endo-peptidase + N-terminal Glu, caspases 1-10, Clostripain, Enterokinase Glutamyl endo-peptidase, GranzymeB, LysC, LysN, Proline-endo-peptidase, Thrombin, Tobacco etch virus protease, Trypsin	NM	NT	N/A	0.41	-1.34	0.00	5.85

ITLIVVVLM	Arg-C proteinase, Asp-N endo-peptidase, Asp-N endopeptidase + N-terminal Glu, caspases 1-Chymotrypsin-high specificity, Glutamyl endo-peptidase, LysC, LysN, Proline-endo-peptidase, Staphylococcal Thrombin, Tobacco etch virus protease, Trypsin	NM	NT	A	0.47	-1.49	0.00	5.88

LFAIIVLAQ	Arg-C proteinase, Asp-N endo-peptidase, Asp-N endo-peptidase + N-terminal Glu, caspases 1-10, Clostripain, Enterokinase, Glutamyl endo-peptidase, GranzymeB, LysC, LysN, Thrombin, Tobacco etch virus protease, Trypsin	NM	NT	A	0.39	-1.33	0.00	5.88

MQPSKTNGL	Arg-C proteinase, Asp-N endo-peptidase, Asp-N endo-peptidase + N-terminal Glu, aspases 1-10, Clostripain, Enterokinase, Glutamyl endo-peptidase, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease	NM	NT	A	-0.22	0.02	1.00	9.11

YSKNRTRTS	Arg-C proteinase, Asp-N endo-peptidase, Asp-N endopeptidase + N-terminal Glu, caspases 1-10, Enterokinase, Glutamyl endo-peptidase, GranzymeB, Neutrophil elastase, Pepsin (pH1.3), Thrombin, Tobacco etch virus protease	NM	NT	N/A	-0.68	0.74	3.00	11.01

IVVVLMQPS	Arg-C proteinase, Asp-N endopeptidase, Asp-N endopeptidase + N-terminal Glu, caspases 1-10, Clostripain, Enterokinase, Glutamyl endo-Proline-endopeptidase, peptidase I, Thrombin, Tobacco etch virus protease, Trypsin	NM	NT	A	0.24	-0.99	0.00	5.88

FAIIVLAQN	Arg-C proteinase, Asp-N endo-peptidase, Asp-N endo-peptidase + N-terminal Glu, caspases 1-10, Enterokinase, Glutamyl endo-peptidase, GranzymeB, Staphylococcal peptidase I, Thrombin, Tobacco etch virus protease, Trypsin	NM	NT	A	0.26	-1.11	0.00	5.88

ILFAIIVLA	Arg-C proteinase, Asp-N endo-peptidase,Asp-N endopeptidase + N-terminal Glu, caspases 1-10, Clostripain, Enterokinase, Glutamyl endo-peptidase, Thrombin, Tobacco etch virus protease, Trypsin N-terminal Glu, caspases 1-10,	NM	NT	A	0.54	-1.56	0.00	5.88

ITLIVVVLM	Arg-C proteinase, Asp-N endo-peptidase, Asp-N endopeptidase + N-terminal Glu, caspases 1-10, Clostripain, Enterokinase, Glutamyl endo-peptidase, GranzymeB, LysC, LysN, Proline-endo-peptidase, Staphylococcal Chymotrypsin-high specificity	NM	NT	A	0.47	-1.49	0.00	5.88

NA: NON-ALLERGENA

A: ALLERGEN

NM: NO MUTATION

NT: NON-TOXIN