An integrated Approach for the identification of Vaccine and Drug Targets against Multi-Drug Resistant *Yersinia enterocolitica*

ومن یؤت الحکمة
فقد اوتی خیرا کثیرا

QUAID-I-AZAM UNIVERSITY
ISLAMABAD

By
Qurat-ul-ain

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-i-Azam University

Islamabad, Pakistan

2017

An integrated Approach for the identification of Vaccine and Drug Targets against Multi-Drug Resistant *Yersinia enterocolitica*

ومن يُؤتَ الحكمة
فقد اُوتیَ خیرًا کثیرًا

QUAID-I-AZAM UNIVERSITY
ISLAMABAD

A Thesis submitted in partial fulfilment of the requirements for the degree of **Master of Philosophy in Bioinformatics at Quaid-i-Azam University, Islamabad.**

By
Qurat-ul-ain

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-i-Azam University Islamabad, Pakistan Islamabad, Pakistan

2017

CERTIFICATE

This thesis submitted by **Miss Qurat ul Ain** from National Centre for Bioinformatics, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad, Pakistan, is accepted in its present form as satisfying the thesis requirement for the Degree of Master of Philosophy in Bioinformatics.

Internal Examiner:____________________
Dr. Syed Sikander Azam
Assistant Professor & Supervisor
Quaid-i-Azam University Islamabad.

External Examiner:____________________
Dr. Muhammad Saeed
Assistant Professor
Department of Bio-Science
COMSATS Institute of Information Technology
Islamabad

Chairman:____________________
Dr. Sajid Rashid
Associate Professor
National Centre for Bioinformatics
Quaid-i-Azam University Islamabad

Date: September 18,2017

DECLARATION

The work reported in this thesis was carried out by Qurat-ul-ain and I hereby declare that the title of thesis, "computational identification of potential drug and vaccine target against *Yersinia enterocolitica*" and the contents of thesis are product of my own research and no part has been copied from any published source (except the references, standard mathematical or genetic models /equations /formulas /protocols etc.). I further declare that this work has not been submitted for award of any other degree /diploma. The University may take action if the information provided is found inaccurate at any stage.

Signature of Scholar

Dedication

This thesis is dedicated to: The sake of Allah, my Creator and my Master, My great teacher and messenger, Mohammed (May Allah bless and grant him), who taught us the purpose of life, My homeland, the warmest womb; My great parents, who never stop giving of themselves in countless ways, My beloved brothers and sister; who stands by me when things look bleak, My friends who encourage and support me, All the people in my life who touch my heart, I dedicate this research.

ACKNOWLEDGEMENTS

In the Name of Allah, the Most Merciful, the Most Compassionate all praise be to Allah, the Lord of the worlds; and prayers and peace be upon Mohamed His servant and messenger.

First and foremost, I must acknowledge my limitless thanks to Allah, the Ever-Magnificent; the Ever-Thankful, for His help and bless. I am totally sure that this work would have never become truth, without His guidance.

I owe a deep debt of gratitude to our university for giving us an opportunity to complete this work.

I am grateful to some people, who worked hard with me from the beginning till the completion of the present research particularly my supervisor Dr. Syed Sikander Azam, who has been always generous during all phases of the research. His encouraging words and useful insights elevated my spirit to resolve complicated problems.

I would like to thank Dr. Sajid Rashid, the head of department. He is an intellectual, hardworking personality whose efforts in establishing the repute and quality education at National Center for Bioinformatics are commendable.

I am grateful to Sajjad Ahmad, and Saad Raza who had been extraordinarily helpful throughout my research tenure.

I would like to take this opportunity to say warm thanks to all my beloved friends, my colleagues, who have been so supportive along the way of doing my thesis.

I also would like to express my wholehearted thanks to my family for their generous support they provided me throughout my entire life and particularly through the process of pursuing the MPhil degree. Because of their unconditional love and prayers, I have the chance to complete this thesis.

Qurat-ul-ain

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Microbial Virulence Factor Database	MvirDB
Protein Basic Local Alignment Search Tool	BLAST
Transmembrane Helices; Hidden Markov Model	TMHMM
University of California at San Francisco Chimera Chimera	UCSF
Molecular Operating Environment	MOE
2-Dimentional	2D
3-Dimensional	3D
Assisted Model Building with Energy Refinement	AMBER
Basic Local Alignment Search Tool	BLAST
Beta Factor	B Factor
Canonical Ensemble	NVT
Cluster Database at High Identity with Tolerance	CD-HIT
Database of Essential Genes	DEG
Discovery Studio Visualizer	DS Visualizer
Expectation-Value	E-Value
Expert Protein Analysis System	ExPASy
General Amber force field	GAFF
Genetic Optimization for Ligand Docking	GOLD
Grand Canonical Ensemble	µVT
Iterative TASSER	I-TASSER
KEGG Automatic Annotation Server	KAAS
Kyoto Encyclopedia of Genes and Genome	KEGG
Molecular Dynamics	MD
Nanoseconds	NS
Nuclear Magnetic Resonance	NMR

Pyridoxal 5′-phosphate Synthase	Pdxj
Outer Membrane Protein C, porin	meoA
Outer Membrane Protein	ompC2
Nuclear Magnetic Resonance	NMR
Picoseconds	ps
Process TRAJectory	PTRAJ
Protein Data Bank	PDB
Protein Structure Analysis-web	ProSA-web
Radius of Gyration	Rg
Root Mean Square Deviation	RMSD
Root Mean Square Fluctuation	RMSF
Simulated Annealing with NMR Derived Energy Restraints	SANDER
Sub Cellular Localization predictor	CELLO
Support Vector Machine	SVM
The Research Collaboratory for Structural Bioinformatics	RCSB
Tripos Force Field	TFF
Universal Protein Resource Knowledgebase	UniProtKB
Visual Molecular Dynamics	VMD

LIST OF FIGURES

LIST OF TABLES

Abstract

Yersinia enterocolitica is a Gram-negative bacillus-shaped bacterium, belonging to the family Enterobacteriaceae and it is the most common bacteriological cause of gastrointestinal disease in humans. Statistically, every year the pathogen accounts for 117,000 illnesses, 640 hospitalizations, and 35 deaths in the United States. It is causative agent of opportunistic infections like enterocolitis, acute diarrhea, terminal ileitis, mesenteric lymphadenitis, and pseudo appendicitis but, if it spreads systemically, can also result in fatal sepsis. The associated mortality rate of the pathogen is 50% and is virtually resistant to penicillin G, ampicillin, and cephalotin. The virulence of the pathogen is due to the presence of a highly conserved 70-kb virulence plasmid, termed pYV/pCD and certain chromosomal genes. Multi-drug resistance behaviors of this emerging pathogen, initiate the research towards the identification of novel drugs and vaccine target. The prime focus of this study is the utilization of *in silico* subtractive genomics approach, and molecular docking for identifying and inhibiting a potential therapeutic target crucial for pathogen *Y.enterocolitica.*

The development of new and effective therapeutic procedures is urgently needed to counter the multidrug-resistant phenotypes imposed by the said pathogen. Based on subtractive reverse vaccinology approach, we have successfully predicted novel antigenic peptide vaccine candidates against *Y. enterocolitica.* The pipeline revealed two proteins; meoA and ompC2 as promising vaccine targets. Protein-protein interactions elaborated the involvement of target candidates in the major biological pathway of the pathogen. The predicted 9mer B-cell derived T-cell epitope of proteins are found virulent, antigenic, non-allergic, surface exposed and conserved in all 9 complete sequenced strains of the pathogen. Molecular docking deciphers deep and stable binding of epitopes in the binding pocket of the most predominant allele in human population-the DRB1*0101. These epitopes of target proteins could serve as a base for the development of epitope-driven vaccine against *Y. enterocolitica*

Molecular dynamics (MD) simulation is used to study the dynamic changes in protein under time dependent atmosphere. Pyridoxine 5'-phosphate synthase was selected among all the druggable protein after passing through several filters including subcellular localization, functional annotation and *in silico* subtractive genomic approach. Pyridoxal

5′-phosphate is the active form of vitamin B_6 that acts as an essential, ubiquitous coenzyme in amino acid metabolism. Comparative modelling helped building the structures and then the active site determiation has been carried out. The best protein model was further subjected to molecular docking studies performed using total library of 219 synthetic compounds along with 500 natural compounds to appraise the inhibitor binding potential.

1. INTRODUCTION

1.1 Yersinia enterocolitica

Yersinia enterocolitica is the most common etiological agent of gastrointestinal disease in many developed and developing countries. Amongst those, yersiniosis due to infection with the bacterium Yersinia *enterocolitica* is the frequently reported zoonotic gastrointestinal disease after campylobacteriosis and salmonellosis in many countries, especially in temperate zones (Masignani *et al.*, 2007). It is a member of the genus Yersinia, which encompasses a heterogeneous collection of facultatively anaerobic bacteria, belongs to the family Enterobacteriaceae. Among eleven species within this genus (Wauters *et al.*, 1988), only three, *Y. pestis*, *Y. pseudotuberculosis*, and *Y. enterocolitica* are regarded as pathogenic for humans. It is motile at temperatures of 22–29°C, while it is nonmotile at normal human body temperature (Kapatral *et al.*, 1996). *Y. enterocolitica* causes almost 117,000 illnesses, 640 hospitalizations, and 35 deaths in the United States every year (Griffin *et al.*, 2014). Children are infected more often than adults, and infection is more common in the winter.

Antimicrobial medical care is suggested in almost all *Yersinia* infections with slight exclusion, the main classes of antibiotics are applied to cope up infections with following bacteria include cephalosporins, penicillins, fluoroquinolones and tetracyclines (Huang *et al.*, 2010). *Y. enterocolitica* is associated with a wide range of clinical and immunological manifestations, responsible for intestinal diseases, including enterocolitis with an inflammatory diarrhea in affected infants and young children; acute terminal ileitis and mesenteric lymphadenitis mimicking appendicitis in older children and young adults, as well as rare extraintestinal manifestations including urinary and respiratory tract infections (empyema), osteoarticular infection (reactive arthritis), erythema nodosum, infected mycotic aneurysm (Miller *et al.*, 1989; Menzies *et al.*, 2010), axillary abscesses (Bottone *et al.*, 1997), and endocarditis (Bottone *et al.*, 1997). Human yersiniosis is primarily acquired through the gastrointestinal tract as a result of ingestion of contaminated foods usually raw or inadequately cooked pork if spreads systemically, can also result in fatal sepsis (Fredriksson *et al.*, 2009).

Within developed countries, incidences of yersiniosis and foodborne outbreaks are appeared to be lower in the United States than that of many European countries (Lee *et al.*, 1990; Trends *et al.*, 2009). *Y. enterocolitica* foodborne outbreaks in the United States have involved young children exposed indirectly during the cleaning and preparation of raw or undercooked pork chitterlings (Butler *et al.*, 1984). In European countries, numbers of reported cases of human in England and Wales are lower than those in other European countries where fewer than 0.1 cases of yersiniosis per 100,000 individuals were reported in the United Kingdom in 2005, in contrast to 12.2 in Finland and 6.8 in Germany, respectively (Kanan *et al.*, 2009). On the other hand, the high prevalence of gastrointestinal illness including fatal cases due to yersiniosis is also observed in many developing countries like Bangladesh (Soltan *et al.*, 2004), Iraq (Okwori *et al.*, 2009), Iran (Cover *et al.*, 1989), and Nigeria (Kapatral *et al.*, 1996), which indicates major underlying food safety problems in low- and middle-income countries. Worldwide, infection with *Y. enterocolitica* occurs most often in infants and young children with common symptoms like fever, abdominal pain, and diarrhea, which is often bloody. Older children and young adults are not out of risk. The predominant symptoms within these age groups have right-sided abdominal pain and fever, sometimes confused with appendicitis. Occasionally, the *Y. enterocolitica* associated complications such as skin rash, joint pains, or spread of bacteria to the bloodstream can also occur. Recently, *Y. enterocolitica* has become of concern worldwide, and foodborne infections have been reported in various countries. *Yersinia enterocolitica* infection stimulates significant morbidity and mortality rates (Toma *et al.*, 1974). The pathogen is increasingly identified in foodborne gastrointestinal infections with an alarming mortality rate of 50% (Kapatral *et al.,* 1996 Toma *et al.*, 1974). The mortality rate is less in case of normal infections but rare cases of sepsis can lead to a very high mortality rate. One hundred seventy five strains (175) of the pathogen, *Yersinia enterocolitica* have been reported, however only fifteen (15) of these strains have been completely sequenced to date. The sequenced data is available on National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/genome/genomes/1041). This study emphasizes on exploration of the genome of the selected strain for identifying potential drug targets. Selected reference strain is (*Yersinia enterocolitica* subsp. enterocolitica 8081), significant information about selected reference is retrieved from NCBI (https://www.ncbi.nlm.nih.gov/) and is depicted in Table 1.1.

Table 1.1. Genomic features of Y. enterocolitica subsp. enterocolitica 8081

Strain	Size(Mb)	GC%	Genome assemblies	Protein
subsp. enterocolitica 8081	4.56813	47.2%	160	4023

1.2 Reverse Vaccinology study on *Y.enterocolitica*

The classical vaccine development approach was first introduced by Louis Pasteur, (Serruto *et al.*, 2006) which led to successful discovery of an effective vaccine against several pathogens. Vaccination strategies are emerging as a viable option to prevent and/or treat multi- or pan- drug-resistant infections. Although, there is currently no licensed vaccine against *Y. enterocolitica* infections. As an emerging and revolutionary vaccine development approach, reverse vaccinology (RV) starts with the prediction of vaccine protein targets by bioinformatics analysis of genome protein-coding sequences. With the initial bioinformatics analysis, RV facilitates rapid vaccine design with less reliance on conventional animal testing and clinical trials. The reverse vaccinology is the most practiced methodology since the discovery of a universal vaccine against serogroup B meningococcal (menB) disease in 2000 and it has also been applied to the development of vaccines against a variety of pathogens such as serogroup B *Neisseria meningitidis* (MenB) (Rappuoli *et al.*, 2000), *Bacillus anthracis* (Pizza *et al.*, 2000), *Streptococcus pneumoniae* (Ariel *et al.*, 2002), *Mycobacterium* *tuberculosis* (Wizemann *et al*., 2001), and *Cryptosporidium hominis* (Betts *et al*., 2002).

Reverse vaccinology involves the use of several *in silico* steps to identify immunogenic antigens from the sequenced genomes/proteomes of a pathogen. It is also important to note that these *in silico* steps/methods provide a comprehensive view of the pathogen genome, essential pathways, virulence determining factors and protein–protein interactions (PPI) among itself and the host proteome.

No licensed vaccine is available for treatment of *Y. enterocolitica* associated infections, further emphasizing on the need of vaccine proteins identification and subsequent validation in animal models. However, the results of *Y. enterocolitica* vaccine development efforts, both preclinical and early clinical, have so far been disappointing. Sterilizing immunity is rarely achieved, even in animal models. Traditional vaccine development offers several limitations comprising hypersensitivity, insufficient attenuation, variable products, less immunogenicity and expensive procedures (Manque *et al*., 2011). In post-genomics era, the use of reverse vaccinology (RV) aid in rapid identification of novel protein candidates circumventing the constrains of conventional vaccinology (Giuliani *et al*., 2006). RV is a set of *in silico* filters that unravel immunogenic antigens in the pathogen proteome and analyze pathogen genomic features, essential and selective pathways, virulence attributes and interacting network of target candidates. Prioritization of potential vaccine candidates relies greatly on a virulent feature of identified proteins, capable of stimulating severe infection pathways in the host (Mora *et al*., 2006). Other reported parameters, which could assist in screening potential vaccine candidates, take account of sub- cellular localization, low molecular weight, less number of transmembrane helices, adhesion probability, allergenicity and antigenic epitope conservation.

Current study incorporates sequential integration of various *in silico* approaches and it is based on the development of a comprehensive computational framework for the identification of putative vaccine targets against *Y.enterocolitica*. It is based on subtractive proteomics to prioritize novel antigenic epitopes by integrating data from biological networks and databases. Essential proteins were first identified, followed by confirming their roles in virulence and immune evoking pathways. Subsequently, non-human homologous proteins were scanned for those epitopes having the ability to bind with both B-cells and T-cells. The genetically invariable epitopes identified here can be further subjected to in vivo testing. Following the proposed pipeline, candidate antigens can easily be identified which will assist in the development of protective immunogens against various other pathogens

1.3 Drug Designing study on *Y.enterocolitica*

In the current era, the search for novel drug targets against ever increasing pathogens with resilient features in their genomes has replaced traditional drug discovery and implied the combination of computational tools and integrated data 'omics' such as genomics, proteomics and metabolomics. This strategy is called computer aided drug designing (CADD), which profoundly reduces the number of compounds to be screened without ever compromising the quality of drug target with added benefits of low cost and less time expenditure. Our study is a kind of such approach which revolves around the subtraction of proteome of *Y.enterocolitica* through logical steps which were completed via computational tools and methods. The complete proteome of a reference strain (subsp. enterocolitica 8081) was retrieved through Uniprot (http://www.uniprot.org/). Fully retrieved proteome was then pass through streamline of computational methods which ensured the separation of proteins that are essential for the survival of bacteria, have reduced homology to the human proteins and are part of metabolic pathways unique to the pathogens. These methods include removal of redundant proteins by CD-HIT (Cluster Database at High Identity with Tolerance) (Huang *et al.*, 2010). Debunking of the non-homologues through Blastp (Protein-Protein Basic Local Alignment Search Tool) (Altschul *et al.*, 1990), retrieval of essential proteins by DEG (Database of essential genes) (Zhang *et al.*, 2008), functional annotation, functional family prediction and identification of unique pathways by SVMProt (Garg *et al.*, 2008) and KEGG (Kyoto Encyclopedia of Genes and Genomes) (Ogata *et al.*, 1999). The above mentioned criteria for the selection of drug target would not only enhance the efficacy of drug and at the same time also improve the therapeutic index by avoiding non-specific protein-protein interaction.

1.3.1 Molecular Docking

Molecular docking is well-established technique in the field of structural biology and CADD which places ligands into binding cavity of macromolecular structures (Cavasotto *et al.*, 2009). Molecular docking has become an increasingly important tool for drug discovery. Molecular docking studies are used to determine the interaction of two molecules and to find the best orientation of ligand which would form a complex with overall minimum energy. The small molecule, known as ligand usually fits within protein's cavity which is predicted by the search algorithm. These protein cavities become active when come in contact with any external compounds and are thus called as active sites. Molecular docking offers a faster and economic solution in modern drug discovery as compared to traditional experimental techniques as thousands of compounds can be screened and evaluated for their drugaability prospects. There are many online small molecule repositories that maintain vast number of compounds in ready-to-dock formats.

To rank perspective compounds, search algorithm explores different conformation and orientations of ligand in protein. The different ligand possess of ligand in protein. The different ligand poses in complexes are then assessed by a scoring function which estimate the respective binding energies. The efficiency and accuracy of search algorithm and scoring function impart quality to molecular docking software. The search algorithm should be able to sample all possible degrees of freedom while searching conformational space within appropriate time period. The scoring function on other side should be able to recognize between different but closely related conformations, rank them and accurately present the thermodynamics of intermolecular interaction between protein ligand in complexes. Today, we know large set of scoring functions some of which are restricted to specific docking software. However scoring functions can be broadly group in to three classes; 1: empirical scoring functions. 2: force field base scoring functions and 3) knowledge base scoring functions (Morris *et al.*, 2008).

1.3.2 *Molecular Dynamics Simulation*

To understand the physical basis and function of biological macromolecule molecular dynamic simulation was performed. Molecular dynamic simulation is the principal method implied for analytical and theoretical study of molecular system, and assists comprehensive insight into the time dependent behavior and conformational changes of molecular system (Azam *et al.*, 2012). However, catalytic sites, protein folding in particularly peptide structure and dynamic have been investigated extensively using MD simulation (Hansson *et al.*, 2002). Biomolecules have different level of dynamic behavior that can be categorized into three parts, local motion, rigid body motion and large scale motion over varying time scale.

1.3.3 *Statistical Mechanics*

Molecular dynamic simulation is based on statistical mechanics in which microscopic details are converted into more observable macroscopic quantities. MD simulation stores information about microscopic observable such as atomic positions and velocities. Macroscopic details are transformed into macroscopic entities like pressure, energy and temperature by using statistic mechanics.

Statistical mechanics is the branch of physical sciences that studies macroscopic systems from a molecular point of view. The goal is to understand and to predict macroscopic phenomena from the properties of individual molecules making up the system. The system could range from a collection of solvent molecules to a solvated protein-DNA complex. In order to connect the macroscopic system to the microscopic system, time independent statistical averages are often introduced. We start this discussion by introducing a few definitions.

Thermodynamic behavior of a large system by the relating its microscopic behavior to the thermodynamic properties like pressure, volume, energy etc. through the key concept of ensemble (Wilde *et al.*, 1998). An ensemble is stated as the compilation of all microstates of a molecular system which share one or more macroscopic properties (Wiederstein *et al.*, 2007).

Micro canonical ensemble (NVE): This statistical ensemble maintains total energy, E and number of atoms, N in a system to fixed values such that each member of the ensemble exhibits same number of atom.N and total energy.E. The statistical equilibrium demands the system to disallow exchange of energy or particles with its environment to maintain fixed volume, V.

Canonical Ensemble (NVT): A statistical ensemble which corresponds to fixed number of atoms, N but total energy of the system is unknown. The system is in thermal contact with a heat bath at fixed temperature, T. NVT is also an isolated system with fixed volume, V.

Isobaric-Isothermal Ensemble (NPT): This statistical ensemble represents the systems with constant pressure, P and constant temperature, T. Total energy, E and volume, V of the system vary at thermal equilibrium however the total number of atoms, N are kept fixed.

Grand canonical Ensemble (μVT): A statistical ensemble of open systems where only temperature, T, chemical potential, μ and volume, V are fixed.

1.3.4 *Classical Mechanics*

Newton's second law of motion is the pioneer of classical molecular dynamic simulations. It constitutes numerical, gradual resolution of Newton's law of motion. From a knowledge of the force on each atom, it is possible to determine the acceleration of each atom in the system. Integration of the equations of motion then yields a trajectory that describes the positions, velocities and accelerations of the particles as they vary with time. From this trajectory, the average values of properties can be determined. The method is deterministic; once the positions and velocities of each atom are known, the state of the system can be predicted at any time in the future or the past. Molecular dynamics simulations can be time consuming and computationally expensive. However, computers are getting faster and cheaper. Simulations of solvated proteins are calculated up to the nanosecond time scale, however, simulations into the millisecond regime have been reported.

The equation of the second law of motion for the *ith* particle is given by

$$F_i = m_i a_i \tag{1.1}$$

Where F_i is the force exerted on particle i, m_i is the mass of particle i and a_i is the acceleration of particle i. since acceleration is the second derivative of distance r with respect to time t, the above equation can be written as

$$F_i = m_i \frac{d^2 r_i}{dt^2} \tag{1.2}$$

Change in potential energy V of mass m can be calculated as the distance covered by the application of force F

$$F_i = -\frac{dv}{dr_i} \tag{1.3}$$

Combining these two equations yields

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \tag{1.4}$$

Where V is the potential energy of the system. **Newton's** equation of motion can then relate the derivative of the potential energy to the changes in position as a function of time.

1.3.5 *Molecular mechanics*

Molecular mechanics is an extension of classical mechanics, and aim at calculating the energy associated with conformation of molecule. The potential energy of the system of interest is calculated by applying force field.

The following functional abstraction, termed a potential function or force field in chemistry, calculates the molecular system's potential energy (E) in a given conformation as a sum of individual energy terms.

$$E = Ecovalent + Enon\text{−}covalent \tag{1.5}$$

$$Ecovalent = Ebond + Eangle + Edihedral \tag{1.6}$$

$$Enon\text{−}covalent = Eelectrostatics + Evan\ der\ Waals \tag{1.7}$$

The sum of electrostatic and Van der walls forces and the covalent interactions including the bond, angles and dihedrals constitutes the non-covalent interactions. Radical and revolutionary results in field of drug discovery can be achieved by the assistance of molecular dynamics simulation and in having in depth knowledge about dynamic behavior of protein, involved in various diseases (Alonso *et al.*, 2006).

1.4. Radial Distribution Function

The radial distribution function (RDF) defines the probability of finding a particle at a distance r from another tagged particle and the average distance of two atom masks based on an orientation vector. It is denoted by g(r). Radial Distribution Function (RDF) for the docked complex before and after simulation were calculated specifically in order to monitor the conformational variation in the studied system.

1.5 Aims and Objectives of Current Research

The current research aimed to address the importance of opportunistic pathogen, *Yersinia enterocolitica* by following well-designed computational protocol. Purpose of this study is to carry out an efficient *in silico* method for identification of novel potential drug and vaccine target of this pathogen along with the exploration of potent inhibitors for the target. Proposed methodology of the study begins with extraction of reference bacterial proteome. The subtractive proteomic approach ensures the resulting drug and vaccine target to be bacterial-specific, conserved and indispensable for bacterial survival. In the absence of experimental structure, homology modeling offers to generate structural coordinates of selected drug target. The major focus was to produce multiple homology models and select the highest possible quality structure after rigorous model evaluation and refinement. The molecular docking was applied in current study to unveil structural and chemical aspects of interactions involved in inhibiting the bacterial protein. To grasp the mechanistic and dynamic aspects of docked complex system, running molecular dynamics simulation was the final step of proposed methodology. The MD simulation was employed to examine the stability of prospective drug-target complex under longer periods of simulation time and to illustrate the conformational changes that take place during the involvement of protein in dedicated catalytic process.

2. MATERIALS AND METHODS

2.1. System Specification

Comparative and subtractive genomic approaches, based on the strategy that proteins encoded by essential genes of pathogen and non-homologous to the host can be used as drug and vaccine targets (Butt *et al.*, 2012). Such an approach has been effectively used to identify drug and vaccine targets in bacterial species. The computational framework performed for vaccine and drug designing can be divided into different categories. Each of these division implied a diverse set of tools and softwares, which were integrated together to insinuate.

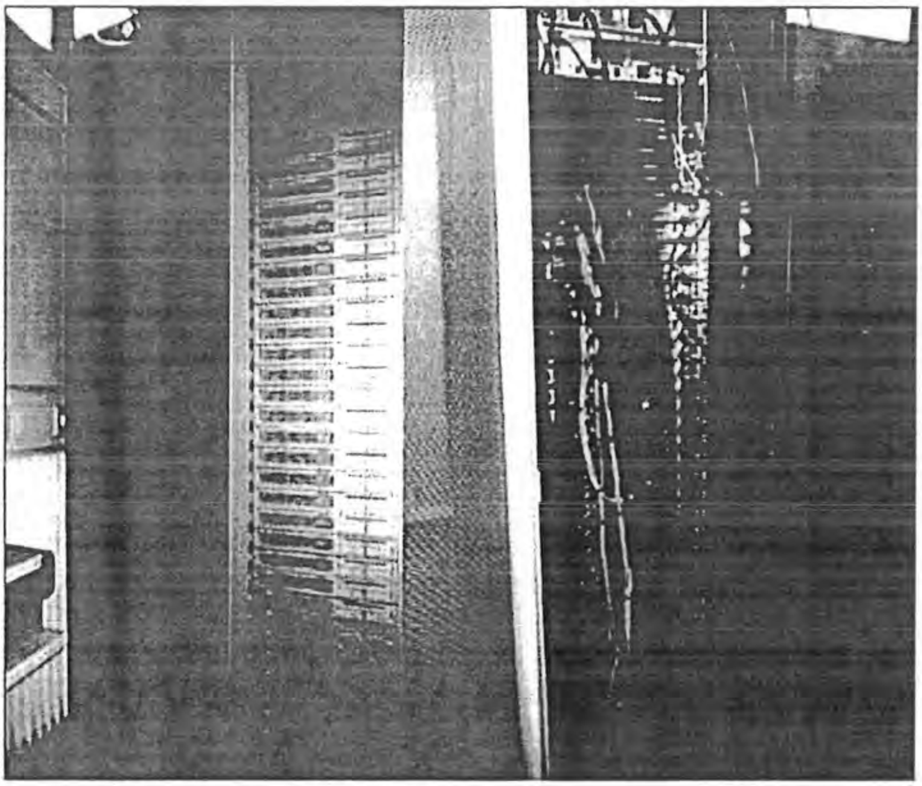


Figure 2.1. *Cluster system used for computer simulations.*

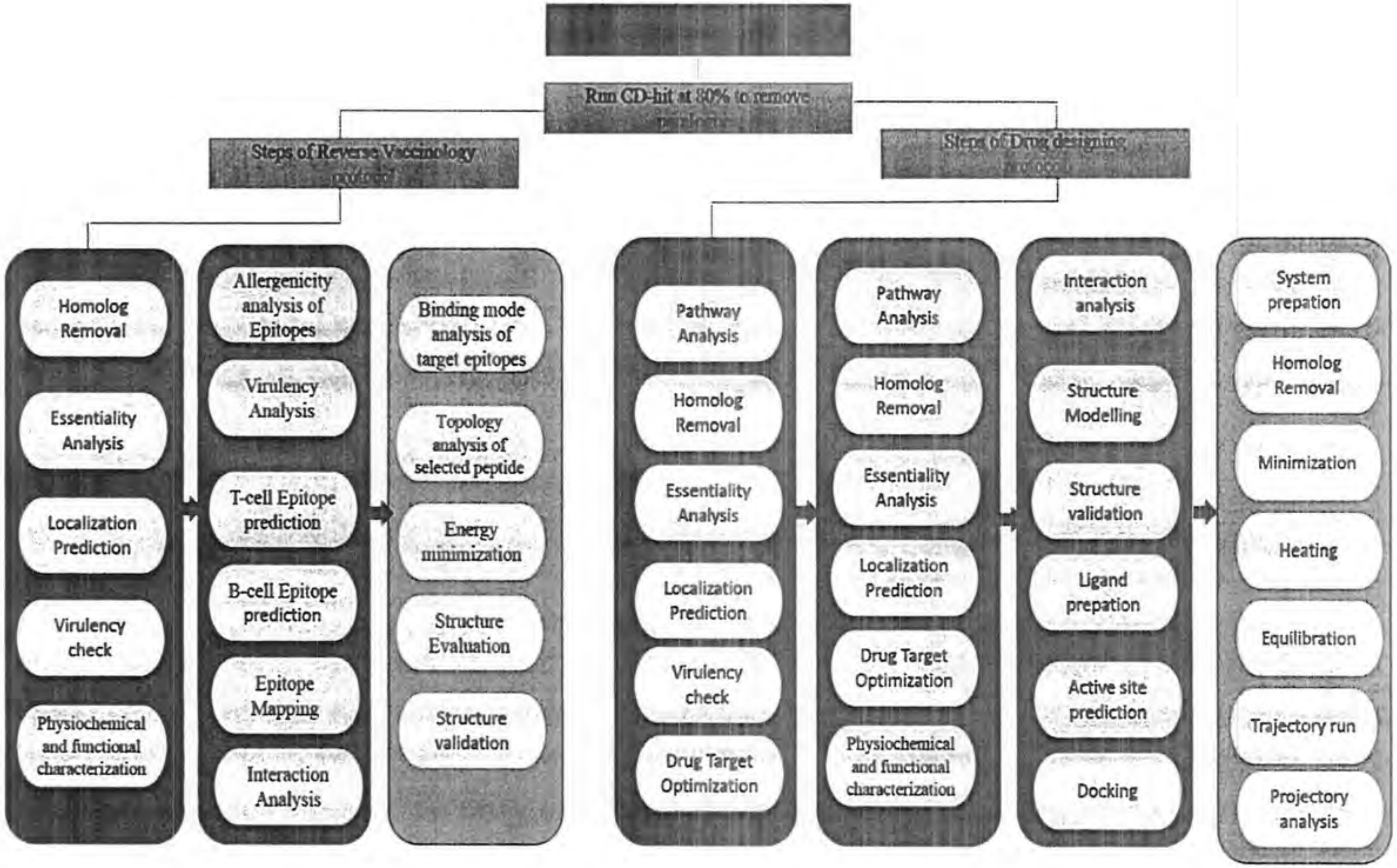


Figure 2.2. *Schematic work flow, highlighting the major steps employed in study.*

2.2. Subtractive proteomics

2.3. Methodology for vaccine target identification

Complete proteome of the reference strain (*Y. enterocolitica* strain 8081) was retrieved from Uniprot (http://www.uniprot.org/) and subjected to the subtractive proteomic pipeline where proteins relevant for vaccine designing were extracted in a step-wise manner. The first step of subtractive proteomics was to remove redundant protein sequences which share an identity of 80%. This was achieved by accessing CD-HIT suitē, (http://weizhongli-lab.org/cd-hit/) which implements hierarchal clustering algorithm to compare biological sequences (Huang *et al*., 2010). The resulting non-redundant set of proteins from redundancy check was evaluated for homology against host (human) proteins. Removal of homologous proteins was imperative while screening vaccine proteins as such proteins result in cross-reactivity with the host proteins generating auto-immune responses. For this, a Perl script was used created by Saad Raza provided by Computational Biology Lab, National Center for Bioinformatics, Quaid-I-Azam University, Pakistan. The non-redundant proteome was aligned against human proteome (Tax id, 9606) using Basic Local Alignment Search Tool (Blast) (Altschul *et al*., 1990) of National Center for Biotechnological Information (NCBI) (https://www.ncbi.nlm.nih.gov/) with threshold expectation value (E-value) set to 10^{-4} and identity cut-off of > 35 % while for alignment and scoring default parameters were employed. Proteins with identity less than the threshold or no corresponding hits were considered as non-homologous proteins and subjected to essential proteome quest. Essential proteins are crucial for the survival of an organism and considered as a foundation to life. Blastp search was performed against DEG (http://tubic.tju.edu.cn/deg/) (Zhang *et al*., 2008) using a Perl script with the threshold E-value of 10^{-10}, bit score of 100 and sequence identity of $\geq 30\%$ was used to extract set of essential proteins.

2.3.1. *Comparative subcellular localization*

The proteins which were designated as non-redundant, human-non-homologous and essential in the previous phase were analyzed for subcellular localization to pool out proteins constituting exoproteome and secretome of the pathogen. The exoproteome and secretome are considered the excellent source of vaccine candidates because of their frequent contact with the biotic and abiotic factor of the extracellular environment. Subcellular localization of the proteins was performed based comparative approach using three online subcellular localization tools: PSORTb (http://www.psort.org/psortb/), CELLO (http://cello.life.nctu.edu.tw/cello2go/alignment.php) (Chen *et al.*, 2006) and CELLO2GO ((http://cello.life.nctu.edu.tw/cello2go/) (Cheng *et al.*, 2014). Proteins predicted as outer membranous and extracellular by all the three tools were only considered and evaluated in subsequent virulent proteins analysis.

2.3.2. *Virulent proteins screening*

Virulent proteins mediate severe signaling pathways in the host cells compare to non-virulent proteins, therefore, can serve as valuable vaccine candidates (Barh *et al.*, 2011). To filter out virulent proteins from extracellular and outer membrane protein dataset, the Virulent Factor Database (VFDB) was accessed (http://www.mgc.ac.cn/VFs/) (Chen *et al.*, 2016). The current version of VFDB covers 30 bacterial genera of medical importance and provide a user-friendly interface to extract virulence factor of a pathogen. These virulent proteins are broadly classified into four categories: defensive, offensive, regulatory and nonspecific virulent proteins. Proteins having Bit score > 100% and identity of ≥ 50% were considered as virulent and brought forth to physicochemical characterization stage.

2.3.3. *Physicochemical characterization of virulent proteins*

Prioritizing proteins based on several physicochemical parameters could assist the selection of proteins suitable for subsequent wet lab investigation. Virulent proteins were prioritized on the basis of molecular weight, number of transmembrane helices and adhesion probability. The molecular weight of proteins is a key parameter while selecting proteins for wet lab evaluation (Naz *et al.*, 2015). Protein usually < 110 kDa in weight are believed as a potential candidate for vaccine designing due to purification ease of such proteins during experimental analysis. Estimation of protein molecular weight was done using molecular weight calculator (http://www.bioinformatics.org/sms/prot_mw.html of expasy server.) of expasy server. Filtered proteins from weight evaluation filter were subjected to TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) and HMMTOP (http://www.enzim.hu/hmmtop/) to compute a number of transmembrane helices of shortlisted proteins (Sonnhammer *et al.*, 1998). Proteins with transmembrane helices < 1 are recognized as attractive targets because such proteins tend to clone and express easily comparing to those harboring a high number of transmembrane helices (Käll *et al.*, 2004). Proteins possessing transmembrane ≤ 1 revealed by both the tools were selected for adhesion analysis. The adhesion of microbial pathogens to host cells is mediated by adhesions proteins, therefore, can act as a worthy character for vaccine candidates (Maritz & Richards 2014). Adhesive proteins prompt the potential of bacteria to bind to biotic and abiotic factors with high efficacy, making subsequent colonization and infection feasible for bacteria. Adhesion analysis was done through SPAAN (ftp://203.195.151.45) to predict adhesions and adhesion like proteins using neural networks (Sachdeva *et al.*, 2004).

2.3.4. *Epitope Mapping*

Mapping epitopes capable of stimulating B-cell immunity and bind to MHC-1 and MHC-II (Major histocompatibility complex) is imperative for vaccine development (Barh *et al.*, 2010). In epitope mapping phase, antigenicity of target proteins was deciphered through VaxiJen (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html) (Doytchinova & Flower 2007). Proteins having an antigenic score > 0.4 were considered as potential antigenic proteins and evaluated further for epitope mapping.

2.3.5. *B-Cell Epitopes*

B-cell epitope mapping for antigenic proteins was achieved using BCPreds (http://ailab.ist.psu.edu/bcpred/predict.html) (Rueckert & Guzmán 2012). Epitopes with score > 0.8 were opted and again subjected to membrane topology analysis via TMHMM (Krogh *et al.*, 2001). B-cell epitopes were further characterized on the basis of antigenicity, hydrophobicity, accessibility, and flexibility with Emini surface accessibility prediction tool (Emini *et al.*, 1985), parker hydrophilicity scale (Parker *et al.*, 1986), kaprplus and Schulz flexibility scale (Karplus & Schulz 1985) and Chou and fashman beta turn prediction tool respectively (Chou 1989). The results of this analysis were cross-referred with the T-cell epitopes and evidently, the common findings were taken as the most probable B-cell epitopes.

2.3.6. *B-cell derived T-cell epitopes*

Surface exposed B-cell epitopes of the prioritized proteins with high antigenicity were taken into consideration to predict B-cell derived T-cell epitope (Barh *et al.*, 2010; Rueckert & Guzman 2012). T-cell epitopes prediction was based on the binding of epitopes to both MHC-I and MHC-II molecules. T-cell epitopes interacting with more than 13 alleles of MHC especially with DRB1*0101 allele and having an inhibitory concentration (IC_{50}) < 100 nm were selected (Guan *et al.*, 2003). In order to predict binding allele for MHC-I and MHC-II, prophred (http://www.imtech.res.in/raghava/propred1/) (Singh & Raghava 2003) and prophred1 (http://www.imtech.res.in/raghava/propred/) (Singh & Raghava 2011) servers were accessed respectively. Virulence and antigenicity of the epitopes were again validated using VirulentPred (http://www.mgc.ac.cn/VFs/) (Garg & Gupta 2008) and Vaxigen (Doytchinova & Flower 2007).

2.3.7. *Protein and epitope allergenicity*

Increase number of vaccines are now observed to be associated with allergenic reactions. To overcome this, scrutinized proteins and their epitopes were examined through AllerTop (http://www.ddg-pharmfac.net/AllerTOP/) (Kadam *et al.*, 2016).

AllerTop is an allergenicity prediction tool based on amino acid descriptors, accounting for residue hydrophobicity, size, abundance, helix and β-strand forming propensities that predict allergic and non-allergic proteins with high accuracy (Dimitrov *et al.*, 2014).

2.3.8. *Proteins interacting network*

Deciphering the interacting network of prioritized epitope proteins could provide a better understanding of the impact of such proteins inhibition on pathogen survival. The interacting network of proteins of interest was unraveled by STRING (http://string-db.org/cgi/network.pl) (Szklarczyk *et al.*, 2014). STRING is a biological database and comprises information of protein-protein interactions (Szklarczyk *et al.*, 2014). STRING database gathers information from a number of sources, including experimental data, computational prediction method, and public text collections. The current version 10.0 harbors information related to about 9.6 million proteins from above 2000 organisms. The targeted proteins were given as input to STRING server and *Y. enterocolitica* 8081 was set as reference strain. The resulting PPI network was analyzed for direct and indirect interactions for targeted proteins. Targeted proteins were also functionally characterized by using STRING database (Szklarczyk *et al.*, 2010) and their role in different pathways reported in Kyoto encyclopedia of genes and genomes (KEGG) database. KEGG is a collection of databases which deals with genomes, biological pathways, diseases, drugs and chemical substances (http://www.genome.jp/kegg/) (Ogata *et al.*, 1999).

2.3.9. *Target proteins structure prediction*

The availability of three-dimensional structure of target proteins was first investigated using blastp search of NCBI against PDB (Bernstein *et al.*, 1977). For proteins with no structural information available, comparative modeling was performed using 5 different online tools; Swiss model (Biasini *et al.*, 2014), CPHmodel (Kushwaha *et al.*, 2014), Phyre2 (Kushwaha *et al.*, 2014), I-Taseer (Wu *et al.*, 2007) and Mod-web (Pieper *et al.*, 2008). For model evaluation, PROSA (Wiederstein & Sippl 2007), VERIFY-3D (Luthy *et al.*, 1992), RAMPAGE (Ramachandran *et al.*, 1966), and

ERRAT (Colovos & Yeates 1993) were used. The best model of each protein was selected and forwarded along the pipeline to energy minimization. This was accomplished using UCSF Chimera (Pettersen *et al.*, 2004) by applying Gasteiger charges with the steepest descent and conjugate gradient was set to 750 steps. The step size was allowed 0.02Å under force field of tripos force field (TFF). The minimized proteins were subjected to pepitope analysis to view the topology of selected epitopes.

2.3.10. *Pepitope Analysis*

To confirm surface exposure of selected epitopes and make sure that these epitopes not get folded in the protein structures, Pepitope analysis (http://pepitope.tau.ac.il/) of protein with their respective epitope was done (Mayrose *et al.*, 2007).

2.3.11. *Molecular Docking*

In last, the epitopes fulfilling all filters of the framework were docked into the binding cavity of "DRB1*0101" (PDB ID "1AQD"), the most common binding allele in the human population. Docking was done by using GalaxyPepDock server (http://galaxy.seoklab.org/pepdock) (Lee *et al.*, 2015). Top complex for each epitope was retrieved and interpreted for binding mode and binding interaction through UCSF Chimera (Pettersen *et al.*, 2004) and ligplot (Laskowski *et al.*, 2011).

2.4. Methodology for Drug target identification

Similar to reverse vaccinology protocol complete proteome of the reference strain (Y. enterocolitica strain 8081) was retrieved from Uniprot (http://www.uniprot.org/) and subjected to first step of subtractive Genomics pipeline which is the removal of redundant protein sequences that share an identity of 80%. This was achieved by accessing CD-HIT suite (http://weizhongli-lab.org/cd-hit/). The allotment of metabolic pathways was the next step of screening procedure. The resultant set of proteins were subjected to KEGG (Kyoto Encyclopedia of Genes and Genomes) database to perform comparative pathway analysis for human and pathogen. KEGG pathway database (Ogata *et al.*, 1999) aided us to determine the metabolic pathway along with its respective function and associated diseases. The next of pipeline was the retrieval of non-homolog proteins. Bacterial proteins homologous to human were removed by the application of Blastp. For this purpose non paralogous protein of the bacteria was aligned against human proteome with threshold expectation value (E-value) set to 10^{-4} and identity cut-off of < 35 %. Blast+ was employed to run NCBI blast on a local server using indigenously prepared Perl script developed at Computational Biology Laboratory, National Center for Bioinformatics. The resultant set of protein were subjected to DEG for essentiality analysis. Essential proteins of *Y.enterocolitica* was retrieved by scrutinizing non-homologous proteins against Database of Essential Genes (DEG) (http://tubic.tju.edu.cn/deg/) (Zhang & Lin 2008) by using a Perl script as mentioned above with the threshold E-value of 10^{-10}, bit score of 100 and sequence identity of ≥ 30%.

2.4.1. *Subcellular Localization prediction*

Subcellular localization of proteins could be used to obtain information about their potential functions. Cellular compartment localization of filtered proteins was determined by online servers. Subcellular localization of the drug targets were carried out by PSORTb (http://www.psort.org/psortb/) (Yu *et al*., 2010), and the results obtained were further validated with CELLO v2.5 (http://cello.life.nctu.edu.tw/cello2go/alignment.php) (Yu *et al*., 2006) and CELLO2GO (http://cello.life.nctu.edu.tw/cello2go/) (Yu *et al*., 2014).

2.4.2. *Virulent Factor Analysis*

To filter out virulent proteins from cytoplasmic protein dataset, the Virulent Factor Database (VFDB) was accessed (http://www.mgc.ac.cn/VFs/) (Chen *et al*., 2016). The current version of VFDB covers 30 bacterial genera of medically importance and provide a user-friendly interface to extract virulence factor of a pathogen. These virulent proteins are broadly classified into four categories: defensive, offensive, regulatory and nonspecific virulent proteins. Proteins having Bit score > 100% and identity of ≥ 50% were considered as potential vaccine targets.

2.4.3. *Physicochemical characterization of virulent proteins*

In order to characterize the virulent proteins as putative drug targets their distribution within the bacterial cell was determined using online prediction server. Molecular weight, theoretical PI, atomic composition, instability index, and grand average of hydropathicity (GRAVY) were predicted through ProtParam.

2.4.4. *Qualitative characterization Benchmark*

Prediction of protein function is an important step for its biological significance (Bottone & Mollaret 1977). In our designed pipeline the function of *Y.enterocolitica* was predicted along their classification to respective functional family, using SVM-Prot (Garg & Gupta 2008) and protFun. *Y.enterocolitica* essential proteins interaction and associations were predicted by STRING (http://string-db.org/cgi/network.pl) (Szklarczyk *et al.*, 2014). The 3D structures of selected list of drug targets with no structural information available, comparative modeling was performed using 5 different online tools; Swiss model (Biasini *et al.*, 2014), CPHmodel (Kushwaha *et al.*, 2014), Phyre2 (Kushwaha *et al.*, 2014), I-Taseer (Wu *et al.*, 2007) and Mod web (Pieper *et al.*, 2008). For model evaluation, PROSA (Wiederstein & Sippl 2007), VERIFY 3D (Luthy *et al.*, 1992), RAMPAGE (Ramachandran *et al.*, 1966), and ERRAT (Colovos & Yeates 1993) were used. The best model of each protein was selected and forwarded along the pipeline.

2.4.5. *Druggability of Potential Targets*

Drugability of potential targets is another important therapeutic prioritization filter that define the possibility of being able to modulate a target with a small-molecule drug. The concept of druggability is most often restricted to small molecules (low molecular weight organic substances) (Knox *et al.*, 2011) but also has been extended to include biologic medical products such as therapeutic monoclonal antibodies. Drug Bank (https://www.drugbank.ca/) is an online tool for biochemical .and pharmacological properties of drug and along with this drug targets and their mechanism can also be accessed. Adjacent to this, information about experimental or approved drug their quantitative structure activity relationship (QSAR) and absorption, distribution, metabolism, excertion and toxicity (ADMET) properties are also combined in DrugBank (Knox *et al.*, 2011). To calculate the druggability of potential target DrugBank version 4.2, with default parameters was used and only those proteins were selected which had bit score > 100.

2.5. Molecular Docking

2.6. Preparatory Phase

2.6.1. Active site determination

Appropriate active cleft in protein tertiary structure for small molecule binding must be investigated for its efficient inhibition. Suitable active cleft is characterized based on size, shape, buriedness and the hydrophobic characterization of particular site (Sousa *et al.*, 2013). The binding pocket and active residues within binding sites were predicted through literature search and alignment of pdxJ protein sequences with Blastp (Bernstein *et al.*, 1977).

2.6.2. *Ligand prepation*

Potential inhibitors against pdxJ protein were collected through pubchem. A database of chemical molecules and their activities against biological assays maintained by the National Center for Biotechnology Information (NCBI). It is available at (https://pubchem.ncbi.nlm.nih.gov/). Literature reported inhibitors library and natural inhibitors were also used against target protein for drug discovery.

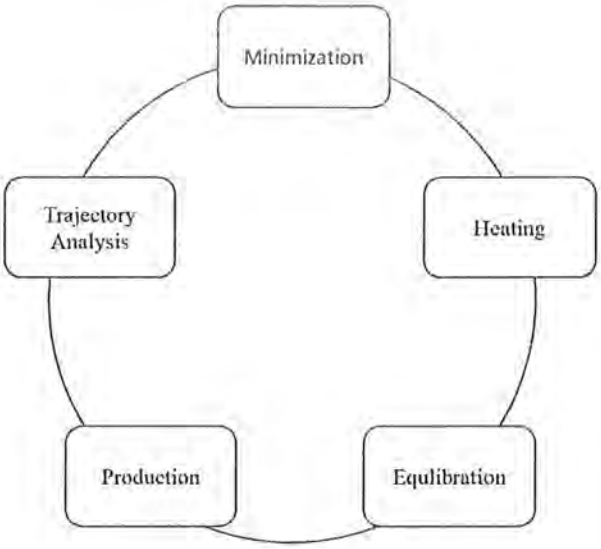
2.7. Molecular Docking Protocol

2.7.1. *Protein-Inhibitor Interaction Analysis*

The molecular interactions between best inhibitor and protein were manually inspected with Chimera and VMD. The interaction diagrams of docked protein was produced with LIGPLOT.

2.8. Molecular Dynamics Simulation

Molecular dynamic simulation of *Y.enterocolitica* targeted protein in complex with the best docked drug-like compound was performed to reveal mechanistic, dynamics and stability details of the target protein with respect to the inhibitor. The complex was first subjected to initial system preparation phase using AMBER 12 software. For ligand processing, antechamber program of AMBER was used while general AMBER force field (GAFF) was employed as a force field. LEap module of AMBER was used to record topology files for protein and inhibitor while MPI and sander was employed to carry out subsequent processing. The system electrostatically neutralized by adding 12 Na+ ions and placed in three-point transferable intermolecular potential (TIP3P) water box of 12 Å size. The system then subjected to a series of seven steps of preprocessing which can be divided into minimization, heating, pre-equilibrium, pressure equilibration and post-equilibrium.



Simulation trajectory analysis

For analysis of results PTRAJ module (process TRAJectory) of AMBER12 was used to produce output files. Following four properties were considered by using PTRAJ and representations by graph were examined in xmgrace (Vaught, 1996).

Root mean square deviation (RMSD)

Root mean square fluctuation (RMF)Radius of gyration (Rg)

β-factor

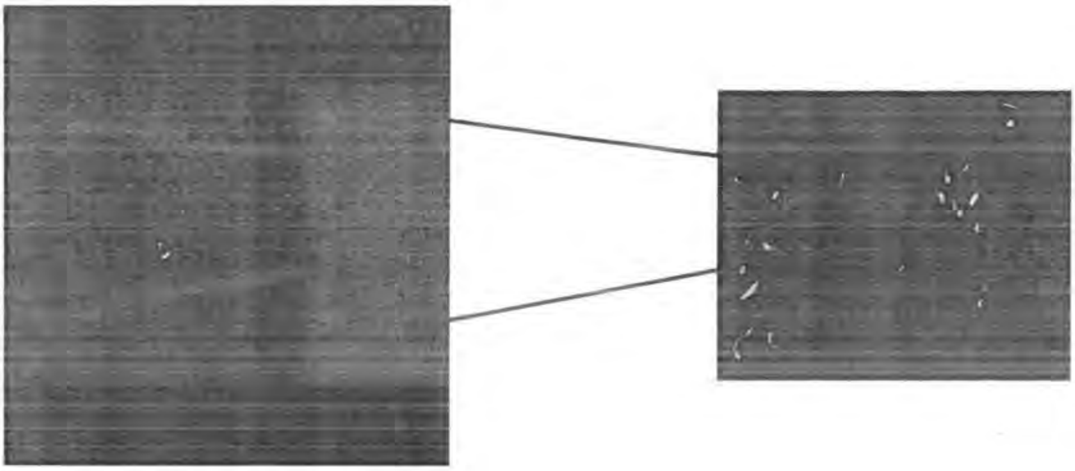


Figure 2.3*. 3D depiction of protein surronded by water molecule.*

2.8.1. ***Root Mean Square Deviation***

The coordinates of alpha carbon (Cα) are generally perceived as reprehensive of the position of an amino acid in three dimensional spaces. RMSD is a measure that allows comparison of relative positions of protein Cα atoms by computation of their averaged distances over a specific time interval (Turner et al., 2007). It is mathematically represented as:

$$RMSD = \frac{\Sigma_{i}^{N}=1d_{i}}{N} \tag{2.1}$$

Where N is the number of compared atoms. Di is the distance between the ith pair of atoms.

2.8.2. *Root Mean Square Fluctuation*

Another important measure of the structural changes in the RMSF that it is calculated for the backbone atoms (N, Cα and C) pf docked target. It represents the root mean square of averaged distance between an atom and its average geometric position in a given set of structural representations (Kuzmanic and Zagrovic, 2010). In context of molecular dynamics, it can be interpreted as to set of positions for an atom that are achieved over a given time scale. Following equation is used to calculate RMSF:

$$RMS\mathrm{F} = \sqrt{\frac{\Sigma_{tj=1}^{T}(x_i(t_j) - x}{T}} \quad (2.2)$$

Where T represents the time interval. Xi represents the position of an atom at a particular time and x represents the averaged position of the atom.

2.8.3. *Beta Factor*

β-factor is a term that is closely linked to the RMSF and measures the spatial displacement of atoms around their mean positions, generated as a consequence of the local vibrational and thermal movements (Kuzmanic and Zagrovic, 2010). Since they measured fluctuations they can be equated in terms of RMSF:

$$\boldsymbol{\beta\ Factor} = \sqrt{\frac{\Sigma_{tj=1}^{T}(x_i(t_j) - x_{((8/3)\,\pi 2)}}{T}} \quad (2.3)$$

2.8.4. *Radius of Gyration*

Radius of gyration is a measure of overall packing quality and density of a structure (Goodfellow, 1990). It is a physical property that can also be experimentally calculated, often through the application of small-angle X-ray scattering (SAXS) (Hong and Lei, 2009). Quantification of the compactness of macromolecular systems was achieved by the implementation of the following equation:

$$Rg = \frac{\Sigma r^2 m}{\Sigma \mathrm{m}} \qquad (2.4)$$

Where N is the total number of atoms, mi is mass of atom „i", ri is position vector of atom„i"and rcm is the centre of mass of molecule under construction.

2.8.5 Radial Distribution Function Analysis

RDFs were calculated specifically to monitor the conformational variation in the studied system.

Radial distribution function (RDF) is a criterion for finding the distribution of atoms, molecules or other species around target.

In the current study, RDFs between amino acids located in the vicinity of pocket and ligand are plotted in order to highlight the conformational variations induced by intermolecular interactions. The radial distribution functions were calculated for the interacting residues crucial for the stability of ligand during the ligand binding process by implementing the following equation.

$$\mathrm{g(r)} = \frac{Pij\ (r)}{<PJ>} = \frac{nij(r)}{<pj>4\pi \mathrm{r}\delta \mathrm{r}} \qquad (2.5)$$

3. RESULTS

3.1. Subtractive Proteomics Approach

This study unveils therapeutic targets that paves a way to open a new horizon for drug and vaccine discovery and could develop treatments for diseases that so far remain intractable. Exploration of potential drug and vaccine targets completed after being successful in various phases of analysis. In the current study, genome of the reference strain of *Y. enterocolitica* was screened and the proteins were shifted at each step on the basis of some user defined threshold.

3.2. Results of Vaccine Target Identification

Due to advances in sequencing technologies and applications of next generation sequencing, a wide-range of genomic data is generated and available in international databases. Till to date, a total of 9 strains of *Y. enterocolitica* are available and all of them were analyzed in the current study. As there is no licensed vaccine available against the pathogen and there is a continuous rise of in resistance strains, identification and subsequent in vitro an in vivo investigation of the proteins for vaccine development is the need of an hour. In this context, the complete proteome of *Y. enterocolitica* strain 8081 was retrieved from UniProtKB, comprising a total of 3957 protein sequences. The proteome CD-hit analysis revealed 3911 proteins as non-paralogues. Screening of non-paralogous sequences was important as paralogous sequences results because of duplication events and often found to have redundant functions. It was therefore considered a healthy exercise to remove such sequence from the complete proteome. Homology analysis based on non-paralogous proteins was aimed to remove proteins shared by both the pathogen and the human host. Removal of homologous proteins was important in the context of vaccine development because of their cross-reactivity with host proteins which can result in autoimmune responses. Homology analysis shortlisted 3324 host non-homologous proteins while 587 homologous proteins were discarded as homologous proteins. The focus was then shifted to essential proteins of the pathogen which is vital for pathogen survival and attractive targets for vaccine development. As such proteins can negatively affect the overall survival of the organism, screening of these proteins was crucial. The analysis pool out total of 1168 proteins as essential while the remaining were excluded from further analysis (**Figure 3.1**).

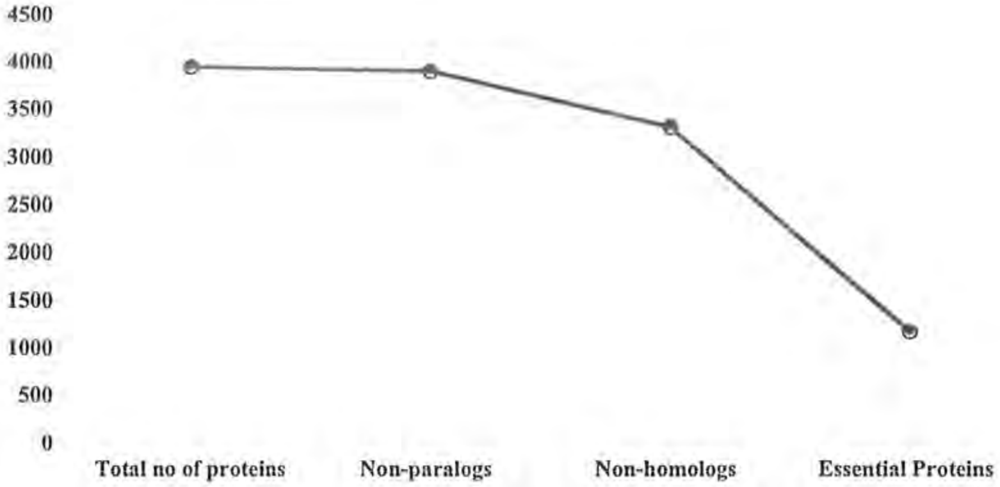


Figure 3.1. *Number of screened proteins obtained at the end of each step of subtractive proteomics.*

3.2.1. Pathogen Exoproteome and Secretome

Exoproteome and Secretome of the pathogen comprise proteins present in the extracellular matrix or outer membrane. Subcellular localization of the essential proteome was an important consideration for identification of potential vaccine candidates. Proteins localized in the outer membrane and extracellular matrix come in frequent contact with the host environment due to their exposed nature. Comparative subcellular localization of essential proteins grouped the majority of the proteins under the cytoplasmic category (**Figure 3.2**). As cytoplasmic proteins are not appropriate targets from vaccine point of view, these proteins were discarded. Psortb analysis based on experimental and computational approaches revealed that 3% of essential proteins are present in the extracellular region while 25% are present in outer membrane region. Similarly, CELLO investigation brought forth 11% of extracellular and 55% of outer membrane proteins. To validate the results of the tools, CELLO2GO demonstrated 4% and 34% of proteins present in extracellular and outer membrane regions. Based on comparative analysis of the results predicted by all the three tools, only 7 proteins present in the extracellular and outer membrane of the pathogen as revealed by all the three tools were considered for virulent protein analysis.

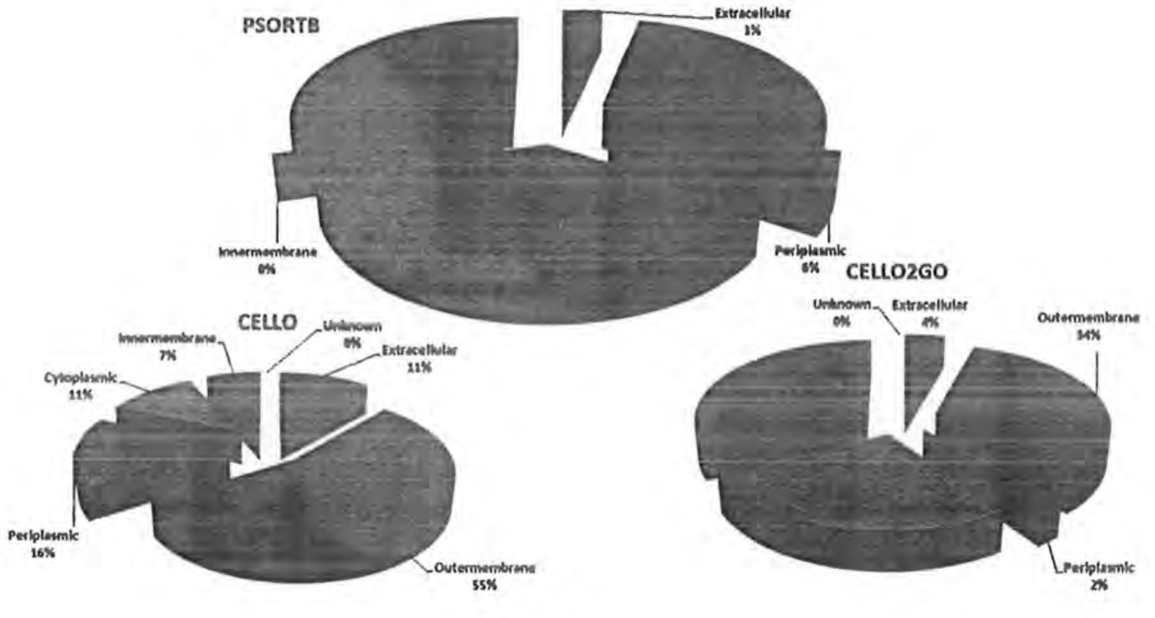


Figure 3.2. *Distribution of proteins on the basis of localization.*

3.2.2. Virulent Proteins Extraction

Virulence factors are the disease-causing proteins associated with microbial pathogenesis and enhance microbial fitness in competing for environment. Immunization of the host with virulent proteins or their combination elicit enhanced protection of the host when exposed to the microbial challenge. In this phase, proteins were examined for their role as virulence factors. Extracellular and outer membranous proteins were blastp against the core database of VFDB to extract data set of virulent proteins. All the 7 proteins were labeled as virulent proteins (**Table 3.1**); HasF is an ABC-transporter outer membrane component which shows 87% identity and 760 bit score with HasA-type hemophore-mediated heme uptake system of *Yersinia pestis*. The protein is involved in molecular functions and directs the movement of molecules including macromolecules, small molecules, and ions out or within a cell, or between cells. MyfC is an outer membrane usher protein which shows an identity of 98 % and 1556 bit score with Mucoid Yersinia factor, Myf fimbriae from *Yersinia enterocolitica* and is involved in the export and assembly of MyfA fimbrial subunits. MeoA is an outer membrane porin protein and shows 62 % identity and 428 bit score with Hek protein from Salmonella enterica subsp. MeoA protein is found in outer membrane of Gram-negative bacteria and involved in molecular functions and catalyze the transportation of small molecule across the membrane. The transmembrane portions of this protein are composed of beta-strands forming a beta-barrel.YE0159 is a putative outer membrane usher protein it shows 50 % identity and 744 bit score with CupA fimbriae from *Pseudomonas aeruginosa*. Putative outer membrane usher protein is involved in transporter activity and enables the movement of substances within or out of the cell. Ompc2 is an Outer membrane protein which shows 56% identity and 388 bit score with Hek from *Salmonella enterica* subsp. enterica serovar Typhimurium strain and it catalyze the transfer of small substances across the membrane. YE2463 is an Outer membrane porin protein that shows 57% identity and 410 bit score with Hek from *Salmonella enterica* and like meoA and YE2463 protein it also catalyzes the transfer of substances across the membrane. OmpA is a Putative outer membrane porin a protein it shows 70% identity and 470 bit score with OmpA from *Escherichia coli* and it is involved in a structural molecular activity which is an action of a molecule toward the structural integrity of a complex or its assembly within or outside a cell.

Table 3.1. *Prioritized virulent proteins of Y. enterocolitica for vaccine designing.*

S.no	Protein	Identity (%)	Bit Score
1	HasF	87%	760
2	MyfC	98%	1556
3	MeoA	62%	428
4	YE0159	50%	744
5	Ompc2	56%	388
6	YE2463	57%	410
7	OmpA	70%	470

3.2.3. Vaccine Proteins Prioritization

List of virulent proteins was further analyzed to get more understanding of the targets and investigate the physicochemical feasibility of the target proteins for experimental validation. Important physicochemical parameters taken into consideration for analysis includes molecular weight estimation, adhesion probability and number of transmembrane helices. As mentioned in the methodology section, molecular weight calculation of proteins is an essential prerequisite from vaccine development point of view, therefore, the inclusion of proteins for further in silico and wet lab studies require the need of removing protein having high molecular weight (Naz *et al.*, 2015). Usually protein < 100 kDa are regarded as suitable candidates for vaccine development as such protein can be easily and efficiently purified. All 7 virulent proteins were found to have molecular weight < 110 kDa (hasF=53.33, myfC= 86.5, porin=40.42, YE0159=81.38, ompC2=40.76, YE2463=39.52, ompA=38.34). After molecular weight analysis, proteins were passed through TMHMM and HMMTOP filters where proteins harboring ≤ 1 were extracted. Proteins with multiple transmembrane helices were discarded because they are not recommended for vaccine development as they are difficult to clone, express, and purify (Kall *et al.*, 2004). All the proteins were found to have 0 transmembrane helices and thus subjected to adhesion probability check. Evaluating adhesion probability of the proteins was critical as adhesion protein mediate the attachment of bacterial pathogen to host tissue and allow bacteria to colonized and cause infections (Sachdeva *et al.*, 2004). Adhesion potential analysis discloses that out of 7 proteins 2 were found to have value <

0.5, thus excluded from the list while the reaming 5 has been characterized as adhesive. The final dataset contained 5 proteins, hasF, meoA, YE0159, ompC2, and YE2463, which satisfied all the filters of the physicochemical filtration phase (**Table 3.2**).

Table 3.2. *List of 5 proteins prioritized on the basis of virulence, subcellular localization, and molecular weight, less number of transmembrane helics, adhesion probability and antigencity.*

Proteins	Cello	Cello2GO	psortB	Molecular weight (kDa)	TMHMM	HMMTOP	Adhesion probability (SPAAN)	Antigencity (VaxiJen >0.4)
hasF	OM	OM	OM	53.33	0	0	0.665	0.56
meoA	OM	OM	OM	40.42	0	0	0.598	0.72
YE0159	OM	OM	OM	81.38	0	0	0.604	0.65
ompC2	OM	OM	OM	40.76	0	0	0.633	0.72
ompC2Y	OM	OM	OM	39.52	0	0	0.601	0.69

3.2.4. Targeted proteins antigenicity

For proteins to act as vaccine proteins need to possess the capability of binding to the products of adaptive immunity such as B-cell and T-cell. The antigenic potential for vaccine proteins is vital, therefore, all the 5 were subjected to VaxiJen server to filter out antigenic proteins. All the 5 proteins were antigenic with the following score, hasF (0.5641), meoA (0.7296), YE0159 (0.6566), OmpC2 (0.7289) and YE2463 (0.6972).

3.2.5. B-Cell Epitope Mapping

After declaring proteins as antigenic, BCpred (http://ailab.ist.psu.edu/bcpred/predict.html) was accessed to map B-cell epitopes for antigenic proteins with epitope length set to 20-mer.The surface exposure of epitopes was confirmed for each targeted protein using TMHMM. BCpred revealed number of antigenic surface exposed B-cell epitopes for each targeted protein with a score higher than 0.8. HasF and meoA protein were found to have 3 B-cell epitopes. YE0159 protein 4 while 2 and 1 B-cell epitopes were mapped for ompC2 and YE2463 protein.

3.2.6. B-Cell derived T-Cell Epitope Mapping

B-cell derived T-cell epitopes were mapped for each consensus B-cell epitope from each targeted protein. B-cell epitopes which bind to a highest number of MHC alleles specifically DRB1*0101 were considered. Epitopes affinity with DRB1*0101 allele produces a good immunogenic response. At this step, two proteins (hasF, YE2463) were excluded as these proteins showed no binding with MHC molecules. For each protein B-cell epitope, shared T-cell epitopes present in both classes of MHC were selected. Following this criteria, epitopes binding to 15 or more alleles were selected. The filter set forth that YGADNFLSQ epitope binds to 23 (MHC-I:20, MHC-II:3), MRPGVSRYN epitope binds to 29 (MHC-I:27, MHC-II:2) and YASSNRTTA epitope binds to 16 (MHC-I:15, MHC-II:1). The selected epitopes were analyzed for their IC_{50} value and epitopes with IC_{50} value < 100nM were selected. Shortlisted epitopes were found to have an IC_{50} value in the following order, (YGADNFLSQ: 36.64), (MRPGVSRYN:481) and (YASSNRTTA:22). At this step putative outer membrane usher protein was excluded from the study as it contains an IC_{50} value above 100 nM. These epitopes were further subjected to VirulentPred and Vaxijen and were found to be antigenic with following score YGADNFLSQ (0.4) and YASSNRTTA (1.53). The epitopes of both proteins were confirmed as antigenic, virulent, and bind to both MHC classes and selected for further studies (**Table 3.3**).

Table 3.3. Final set of prioritized B-cell derived T-cell epitopes against Y. enterocolitica.

Proteins	B-cell derived T-cell epitopes	Total Alleles binding	IC50 value (MHCPred<100 nm)	Viruleny Virulent Pred (>0.5)	Antigency (VaxiJen >0.4)	Allergencity
meoA	YGADNFLSQ	23	36.64	1.06	1.27	Non-allergen
Ompc2	YASSNRTTA	16	22.96	1.06	1.53	Non-allergen

3.2.7. Allergenicity Prediction

An important step in epitope selection for vaccine development is to evaluate its allergenicity. Number of vaccines are now reported to induce allergic responses. To avoid this, the shortlisted epitopes were investigated through AllerTop. Both the epitopes were found to be non-allergen and subjected to conservation analysis.

3.2.8. Epitopes Conservation

Epitopes conservation among fully sequenced strains of pathogens is central for designing broad spectrum vaccines. Epitopes which with complete conservation could be used for broad spectrum vaccine designing while those strains specific conservation could be utilized for strain specific vaccine designing. The two targeted protein sequences from all the 9 strains of the pathogen were retrieved and aligned in CLC sequence viewer. Both the selected epitopes were found completely conserved in 9 strains of the pathogen, thus can be excellent targets for broad-spectrum vaccine development (**Figure 3.3**).

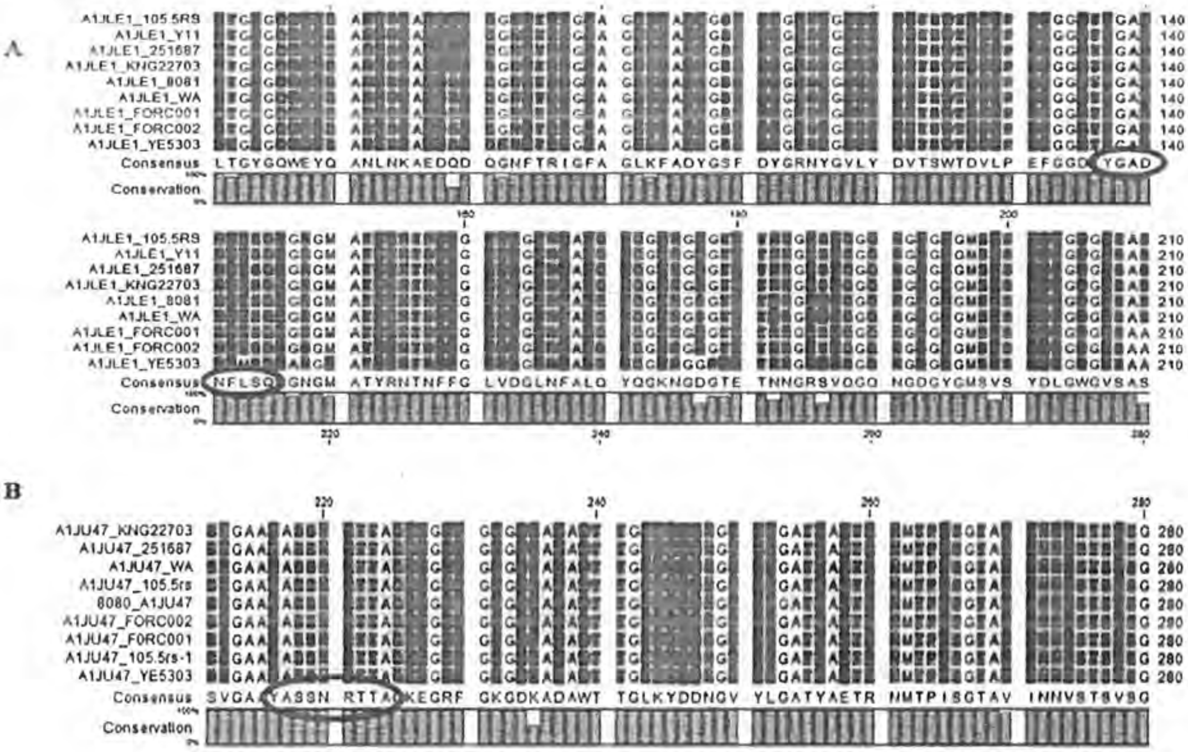


Figure 3.3. Epitope conservation among all sequenced strain for meoA protein (A) and Ompc2 proteins (B) of Y. enterocolitica. The red boxes indicated the conserved 9mer sequence.

3.2.9. Interacting Network of Two Proteins

Interacting partners of a protein could unravel many key signaling pathways and cellular process. Protein-protein interaction of target proteins could provide a better understanding of the impact of target proteins inhibition on overall survival of the pathogen. Similarly, meoA porin is involved in two major pathways, beta-lactam resistance, and two component system. The protein interacts with OmpA which is a major virulent protein and function as a transporter, pores ion channels and structure molecular activity. Other interacting partners include invA, Skp periplasmic chaperone (A protein that allows enteric bacteria to penetrate cultured mammalian cells and aid in cell adhesion), IpoB (Peptidoglycan synthesis regulator and is essential for the function of penicillin-binding protein 1B), BamA (A part of outer membrane protein assembly complex and involved in assembly and insertion of beta-barrel proteins into the outer membrane), YE4066 (function as metal ion binding and involve in peptidase activity), Iptc (Involved in the assembly of lipopolysaccharide (LPS), required for translocation of LPS from the inner membrane to the outer membrane, facilitates the transfer of LPS from the inner membrane to the periplasmic protein LptA and could be a docking site for LptA) and YE0040, YE2851, YE1483, YE0412 proteins (hypothetical proteins belong to Insulinase family protease) **(Figure 3.4A).**

The targeted outer membrane protein (OmpC2) is one major outer membrane protein of *Y. enterocolitica* with high immunogenicity. It belongs to the gram-negative porin family. The protein was found to interact directly and indirectly with several important proteins. An important interacting node is IpoB proteins which are a putative lipoprotein and acts as a regulator of peptidoglycan synthesis and essential for the function of penicillin-binding protein 1B PBP1. Other network proteins includes bamA (an outer membrane protein assembly factor), YaeT (Part of the outer membrane protein assembly complex and is involved in assembly and insertion of beta-barrel proteins into the outer membrane), YE4066 (involve in metal ion binding and peptidase activity), YE0040, YE1483, YE0412 and lptC (Hypothetical proteins associated with assembly of LPS. These proteins are also required for translocation of LPS from the inner membrane to the outer membrane and facilitate the transfer of LPS from the inner membrane to the periplasmic protein. LptD protein along with LptE, carried out the assembly of lipopolysaccharide (LPS) at the surface of the outer membrane **(Figure 3.4B).**

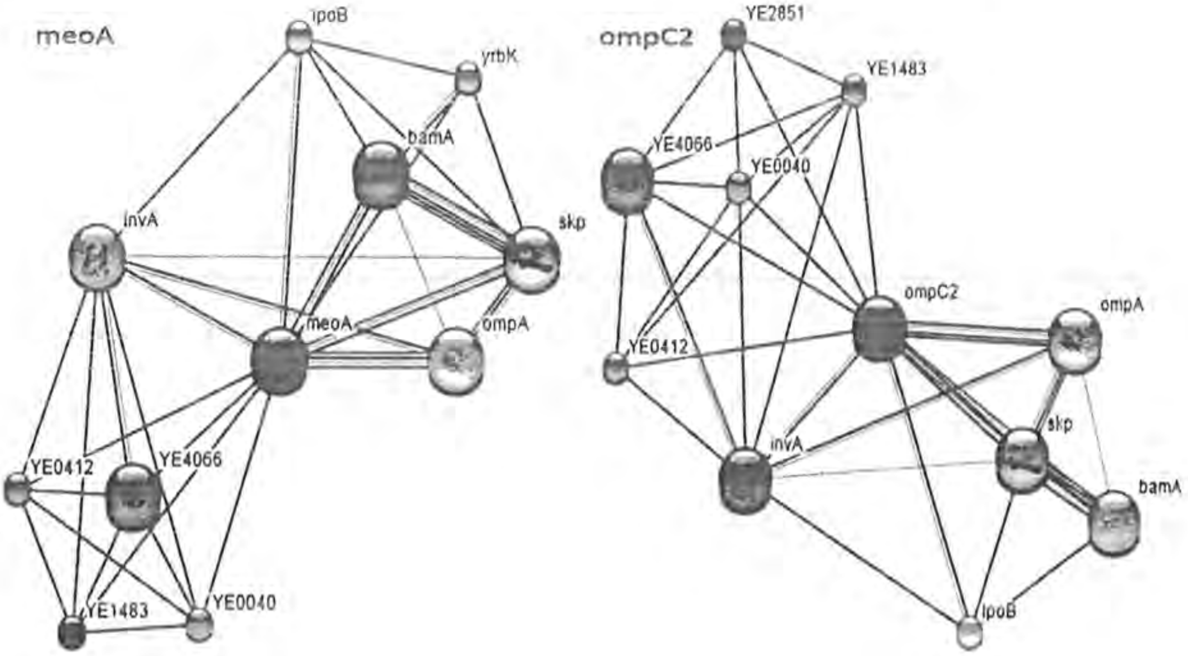


Figure 3.4. *Interactome analysis, highlighting the major interactions of selected proteins with other proteins.*

3.2.10. Comparative Modelling

To view epitope topology on the protein surface, the 3D structure of both the proteins was predicted based on a comparative modeling approach. Models having a low number of residues in the unfavorable region, higher residues in the most favorable region, high verify3D value, a good quality score of Errat and low Z-score were considered as an optimal model. For both proteins, the best model generated by Swiss model was elected due to their high quality score, verify-3D value and least Z-score. In the case of meoA, only one residue were observed in the disallowed region while number of residues in the most favorable region and additionally allowed regions were 320 and 22 respectively. The model structure has high ERRAT quality factor of 86.25, VERIFY-3D score of 78.75 and low Z-score of -2.91. Similarly, OmpC2 protein model was considered as best models as the residue in disallowed region was 3, most favorable region was 319 and in additionally allowed region were 26. The ERRAT quality factor, VERIFY-3D, and Z-score were 76.08, 74.86 and -4.36 respectively. The 3D structure of both modeled proteins is illustrated in **Figure 3.5**.

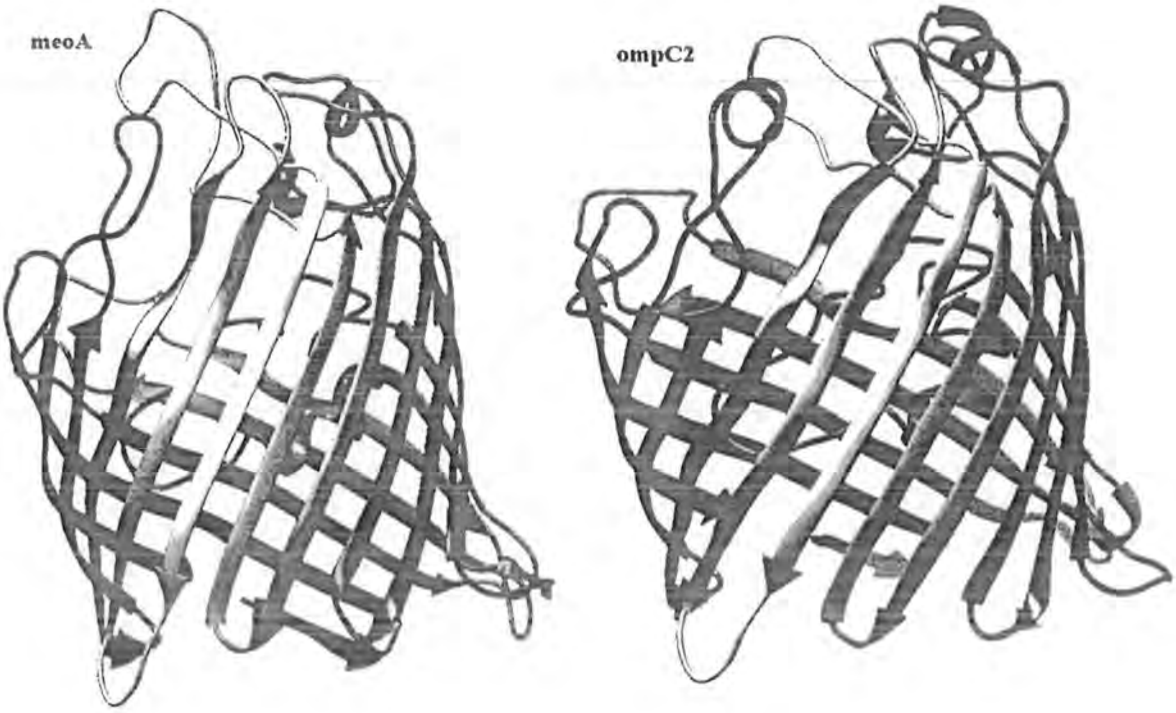


Figure 3.5. *3D depiction of meoA protein and ompc2 protein generated by Swiss-Model.*

3.2.11. Pepitope Analysis

The exo-membrane topology of antigenic, virulent, immunogenic, non-allergen and conserved epitopes was exposed by pepitope server. The analysis showed that the targeted epitopes are surface exposed and not folded in the globular structure of the proteins **Figure 3.6**.

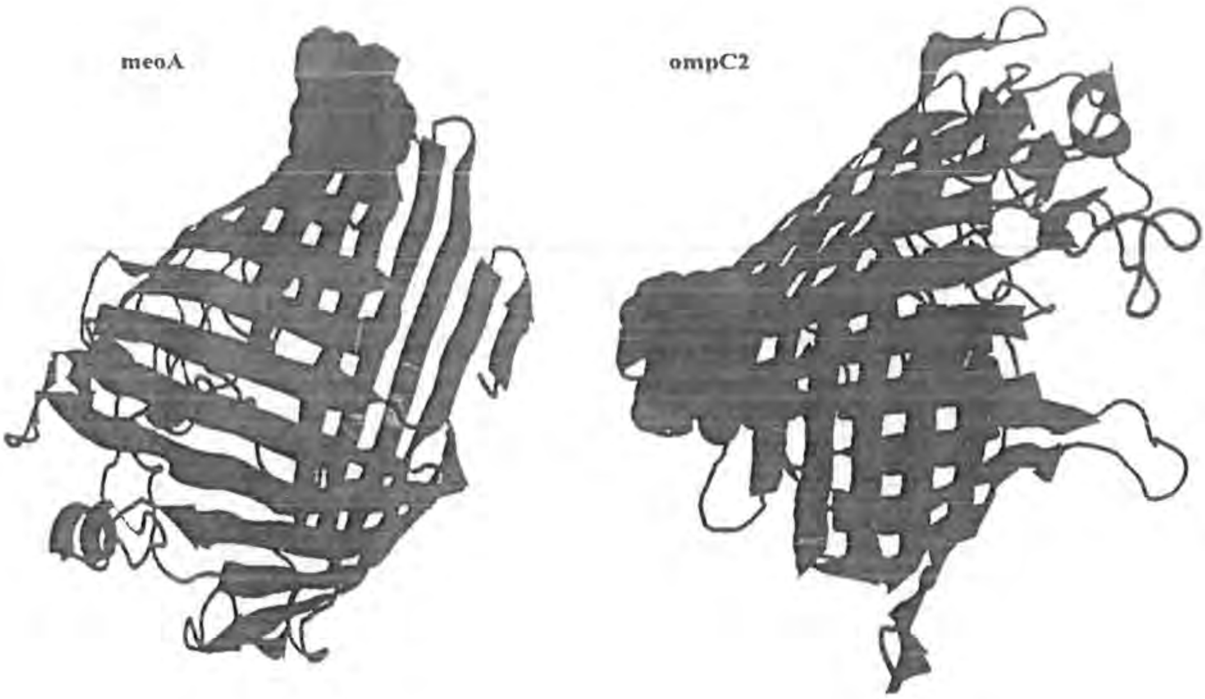
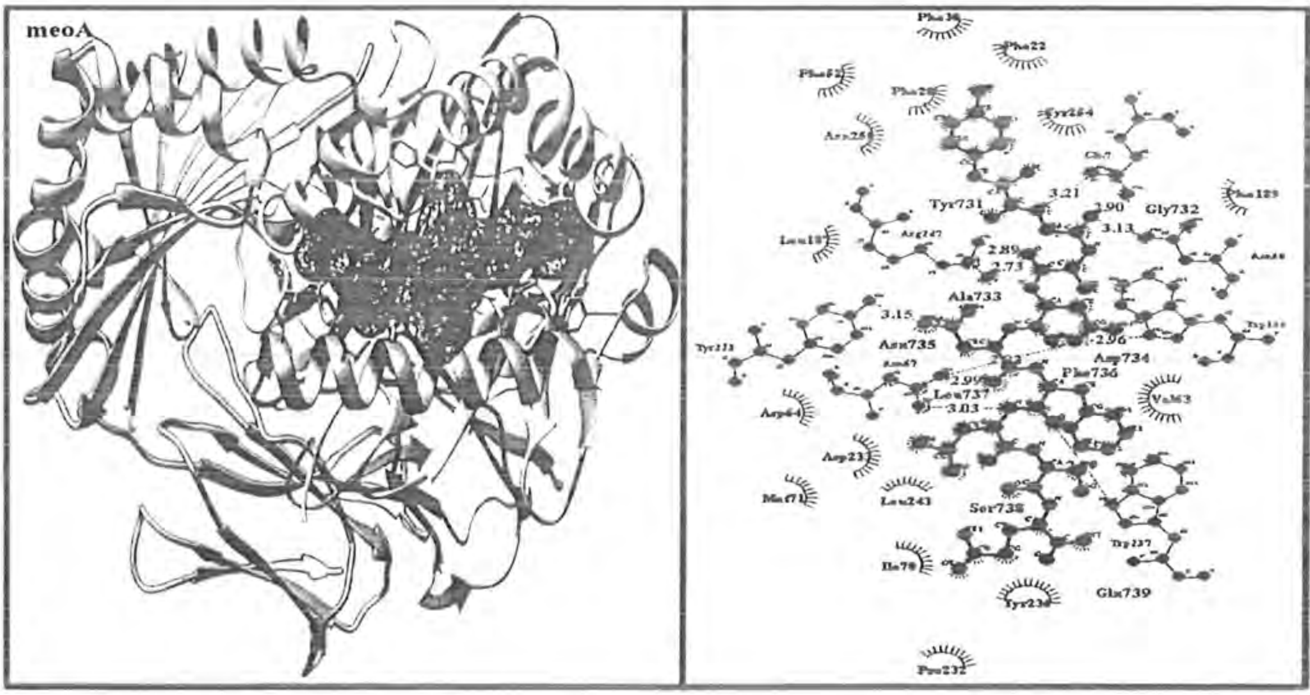


Figure 3.6. Pepitope analysis of both proteins meoA and ompc2 shown in Figure. For both proteins, the antigenic, virulent and non-allergen peptide-based epitopes are shown in red cartoon style present on the surface and doesn't show any folding within a protein while the grey regions represent the proteins.

3.2.12. Molecular Docking Analysis

Both proteins epitope were found deeply bounded into the binding pocket of a DRB1*0101 allele with number of hydrophobic and hydrophilic interactions shown in Fig. In case of meoA protein Ligplus analysis demonstrated that 7 residues of protein (Trp 237 , Trp 185, Asn 60, Gln7, Arg247, Tyr223 and Asn67) were found to form stable hydrogen bonds and 14 hydrophobic interactions were observed between the epitope and DRB1*0101 allele. Similarly, in case of ompC2, 8 residues (Asn67, Glu204, Gln7, Ser51, Asn258, Arg247, Asn60 andTrp237) were found to be involved in forming hydrogen bond with the epitopes while multiple hydrophobic interactions were observed, thus stabilizing the overall complex with the binding mode is **Figure 3.7**.


meoA
Phe36
Phe22
Phe52
Phe26
Tyr254
Asn258
Gln7
Tyr731
3.21
2.90
Gly732
Phe189
3.13
Leu18
Arg247
2.89
2.73
Asn46
Ala733
3.15
Tyr233
Asn735
2.96
Trp158
Asp734
Asn47
2.92
2.99
Phe736
Leu737
3.03
Val63
Asp64
Asp23
Met71
Leu243
Ser738
Trp237
Ile70
Tyr236
Gln739
Pro232

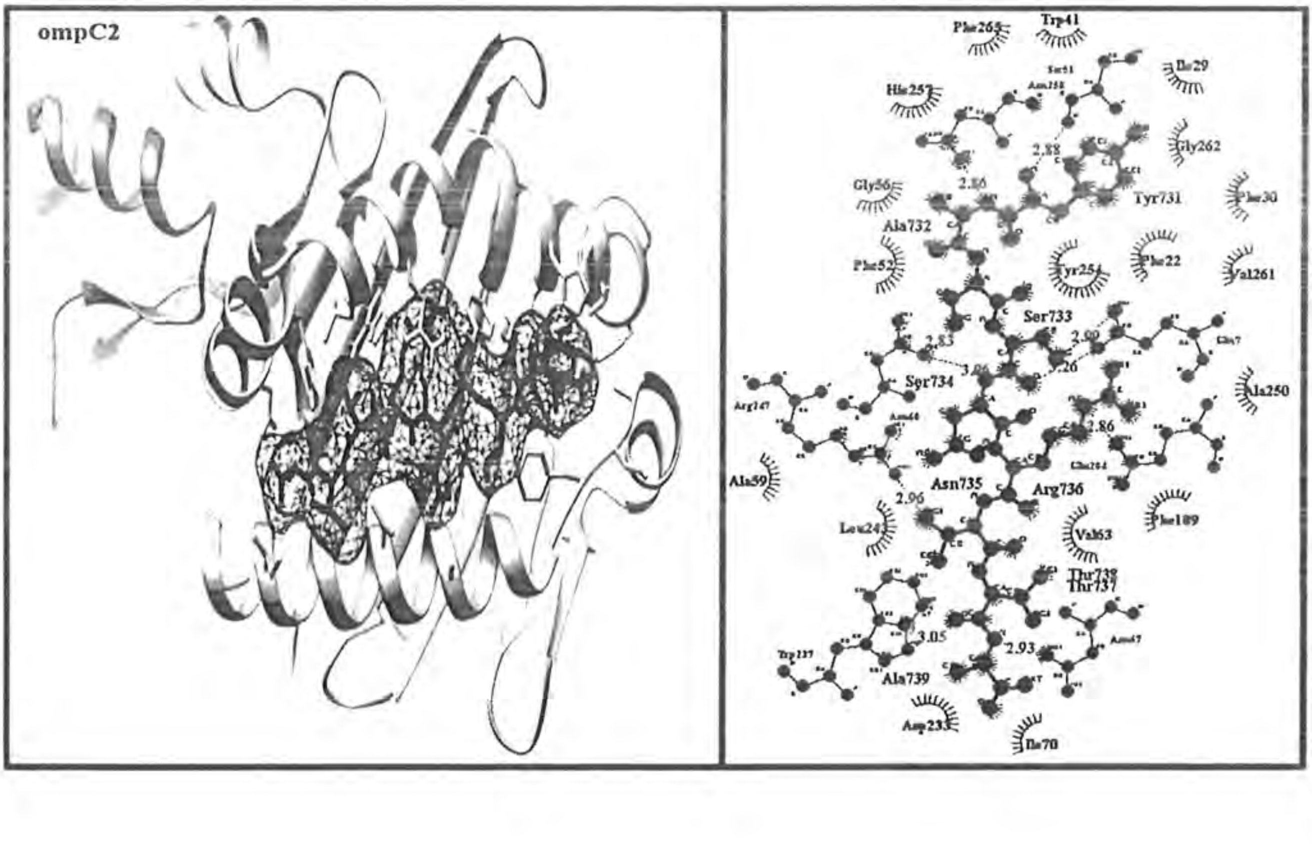

ompC2
Phe265
Trp41
Ile29
Ser51
Asn258
His257
2.88
Gly262
2.86
Gly56
Tyr731
Phe30
Ala732
Phe52
Tyr254
Phe22
Val261
Ser733
2.83
2.99
Gln7
3.06
3.26
Ser734
Ala250
Arg247
Asn46
2.86
Gln254
Ala59
Asn735
Arg736
2.96
Leu243
Val63
Phe189
Thr738
Thr737
3.05
2.93
Trp237
Asn47
Ala739
Asp233
Ile70

Figure 3.7. Docked pose and LIGPLOT illustration of meoA and ompC2 interactions with their surrounding protein residue are demonstrated in figure (a) and (b), respectively.

3.3. Results of Drug Target Identification

3.3.1. *Retrieval of Genome*:

As mentioned above in Reverse Vaccinolgy protocol the complete proteome of *Y. enterocolitca* strain 8081 which is fully sequenced was retrieved from UniProtKB, comprising a total of 3957 protein sequences. After genome retrieval, the first step was to remove the Non-Paralogous proteins from the genome of selected strain. It was performed by CD-HIT application. The proteome CD-hit analysis revealed 3911 proteins as non-paralogues and 46 proteins as paralogs. Set of proteins obtained in first step were then subjected to KAAS for metabolic pathway analysis. KASS identified 414 proteins involved in unique metabolic pathways. Using Perl script, Blastp was performed against human proteome to identify non-homologous proteins and a total of 127 homologous proteins were removed from the proteome leaving only 287 non-homologous proteins in ref strain, respectively. To identify essential proteins in the pathogen which were imperative for their survival a search was performed against the DEG database after the DEG screening, essential proteins identified were 189 and 98 non-essential proteins were left, respectively. The non-essential proteins were not included in the further analysis, due to their insignificance in bacterial survival. An overview of the progressive subtractive genomic screening procedure is shown in **Figure 3.8**.

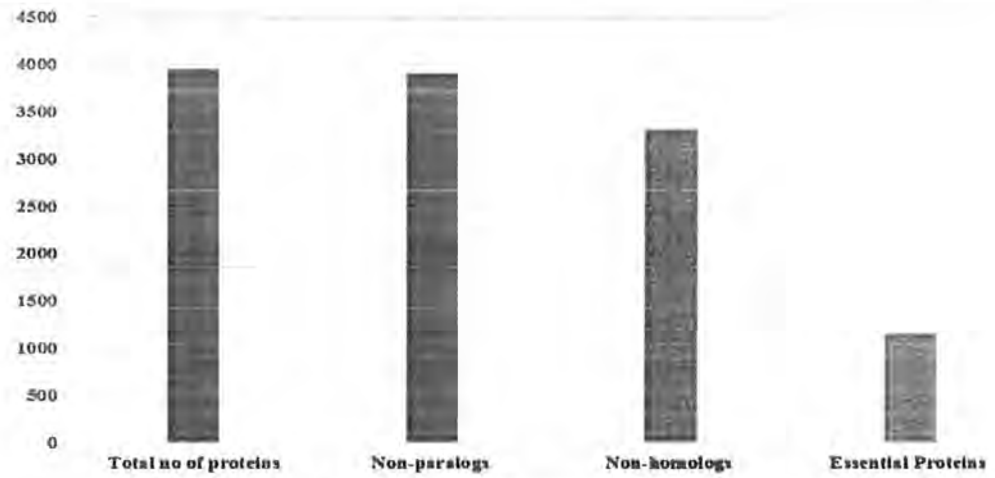


Figure 3.8. Overview of the screened proteins obtained at the end of each step of subtractive genomics.

3.3.2. Subcellular Localization

Drugable targets were then further processed for subcellular localization predictions and 121 cytoplasmic proteins were identified remaining proteins were either periplasmic or membranous. Cytoplasmic proteins were considered to be a potent candidate for being a putative drug target as the cytoplasmic proteins are usually enzymatic in nature, and thus aid bacterial growth.

3.3.3. Virulent Factor Analysis

The proteins were further screened on the basis of experimentally validate pathogenicity analysis to examined their role as virulence factors. In *Y.enterocolitica* virulence proteins were identified from VFDB (virulent factor database). In this phase 23 out of 121 proteins were labeled as virulent proteins.

3.3.4. Druggability Assessment

To ascertain the relation between screened targets and their drug binding ability, DrugBank was used and four druggable targets were identified in *Y.enterocolitica.*

During the screening procedure, many novel targets proteins were also identified for which no hit was scored in Drug Bank.

3.4. Drug Target Selection

Unique pathways that were identified in *Y.enterocolitica* are listed in **Figure 3.9.** The short listed potential drug targets brought forth by applying screening subtractive proteomic approach on *Y.enterocolitica* proteome, hold the information about their subcellular location and metabolic pathway involvement.

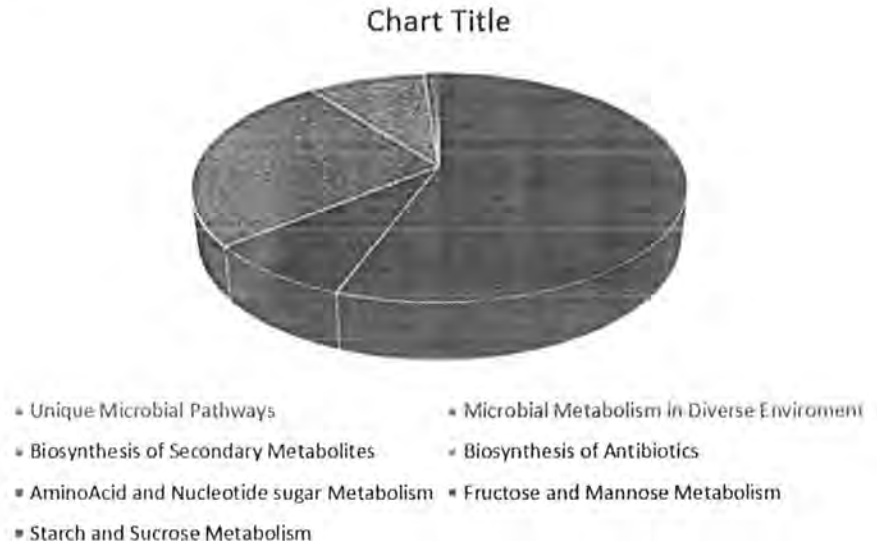


Figure 3.9. ***The representation of bacterial specific essential proteins in major metabolic pathways of Y.enterocolitica.***

3.5. Physicochemical Classification

Selected list of proteins was further evaluated through physiochemical properties to identify possible drug targets. Important physicochemical parameters taken into consideration for analysis includes, Molecular weight, Theoretical PI, Stability and Hydropathicity. The average value of theoretical PI in *Y.enterocolitica* protein was 5.8 that indicate the acidic nature of all selected proteins and the average value of stability index was less than 40 indicating the high stability of proteins even at evaluated temperature conditions. Our 4 selected proteins were found to have molecular weight (36571, 35730, 26209, and 46891). Functionality and physiochemical properties of selected proteins are given in **Table 3.4**.

Table 3.4. *Broad spectrum essentiality analysis on the basis of functionality and physiochemical analysis.*

No.	UniProt ID	Protein length	Structure	Subcellular Localization	Funcionality	TMH	Molecular weight	Theoratical Pi	Instability Index	Hydropathicity
1	HEMH_YERE8	322aa	No	Cytoplasmic	Catalyzes the ferrous insertion into protoporphyrin	0	36571.1	6.72	56.37	-0.274
2	A1JP71_YERE8	340aa	No	Cytoplasmic	Catalyzes the N-acylation of UDP-3-O-glucosamine, involve in biosynthesis of lipid A	0	35730.0	5.79	35.27	0.187
3	A1JKK6_YERE8	243aa	No	Cytoplasmic	Catalyzes the complicated ring closure reaction between the two acyclic compounds 1-deoxy-D-xylulose-5-phosphate (DXP) and 3-amino-2-oxopropyl phosphate (1-amino-acetone-3-phosphate or AAP)	0	26209.2	5.83	46.73	0.104

4	A1JQ97_YERE8	430aa	No	Cytoplasmic	ATPase activity, ATP binding	0	46891.4	5.23	4.54	-0.026

3.6. Homology Modeling

The starting point for Homology Modeling was the structural availability of the selected proteins. Since the experimentally resolved structure was unavailable for all four proteins, comparative model building was carried out. Protein with the best identity and query coverage was selected for model building. Template selected for modeling process was 3F4N, chain A and chain B and 93% identity with 100% query coverage was present between pdxJ target protein and template sequence.

Using the template as a reference, structural models were generated using MODELLER9.14, PHYRE, CPH-model and SWISS-model. A detailed comparison of the stereochemical properties was done to select the best modeled structure (**Table 3.5**). On the basis of physicochemical properties and quality assessment measures, Model generated via SWISS-model was selected for further processing shown in **Figure 3.10**. Alignment between the protein and template illustrating a satisfactory level of conservation is shown in **Figure 3.11**. Besides significant coverage, selected Model showed strong stereochemistry with very few residues in disallowed regions and lowest value of Z score (Table 3.5). Ramachandran plot of the selected model is shown in **Figure 3.12 (a)** where maximum residues are present in the most favored regions. Furthermore, the Z-score of the selected optimum model is plotted in **Figure 3.12 (b).** Superimposed structure of template and the target is shown in **Figure 3.13**.

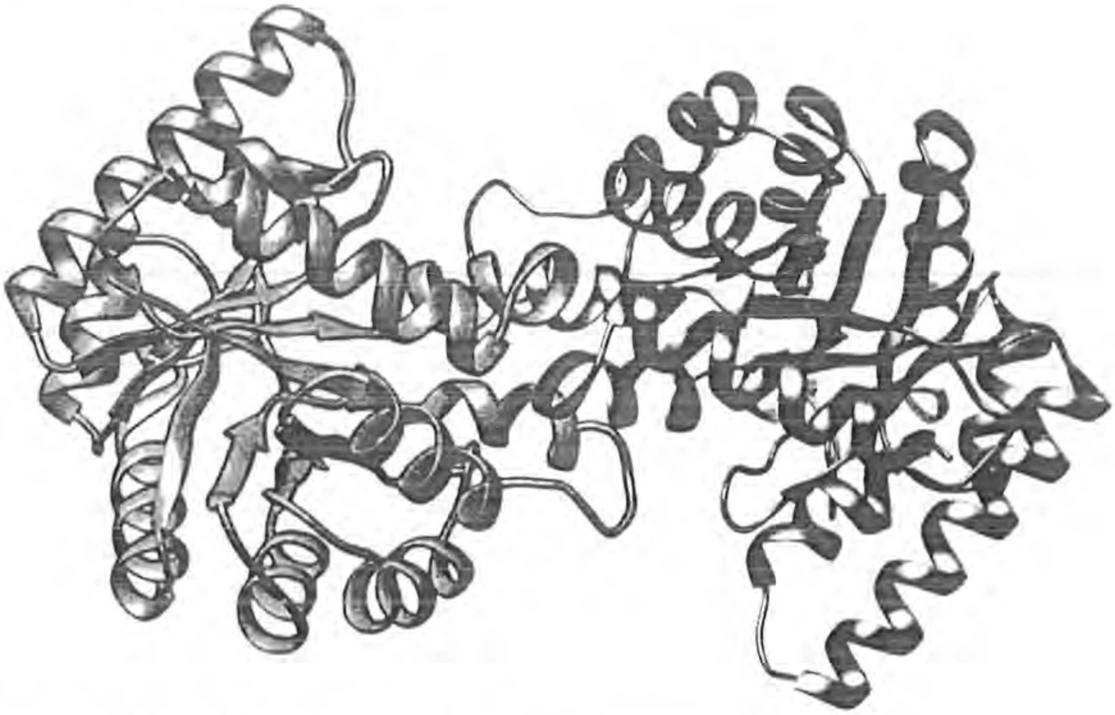
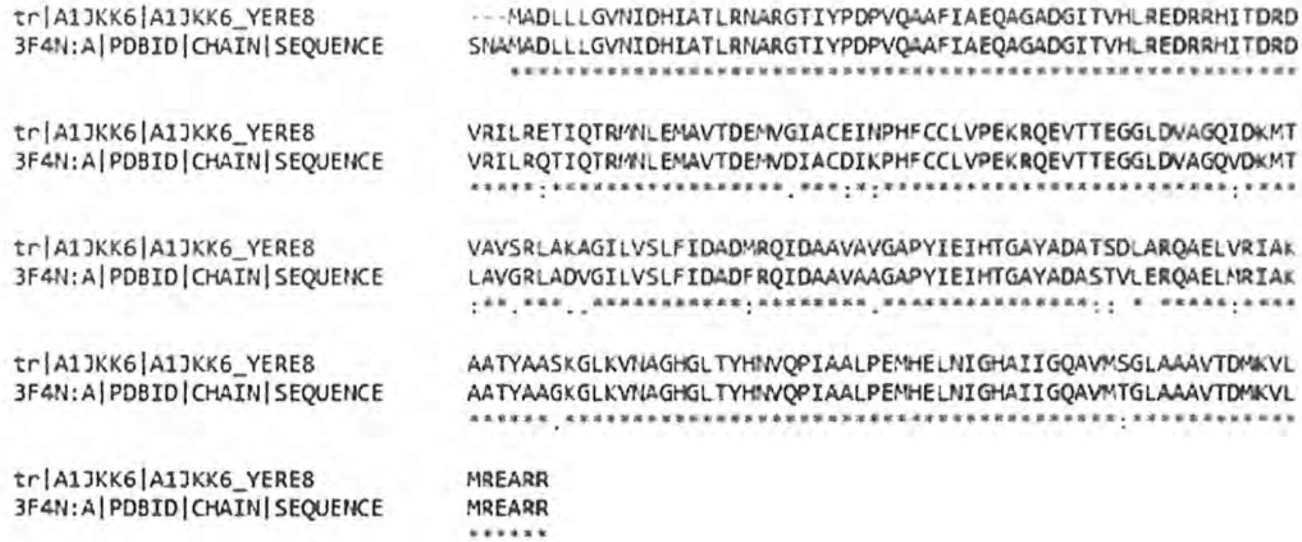


Figure 3.10. Model of pdxJ generated via SWISS-MODEL.



```
tr|A1JKK6|A1JKK6_YERE8              ---MADLLLGVNIDHIATLRNARGTIYPDPVQAAFIAEQAGADGITVHLREDRRHITDRD
3F4N:A|PDBID|CHAIN|SEQUENCE         SNAMADLLLGVNIDHIATLRNARGTIYPDPVQAAFIAEQAGADGITVHLREDRRHITDRD
                                       *********************************************************

tr|A1JKK6|A1JKK6_YERE8              VRILRETIQTRMNLEMAVTDEMVGIACEINPHFCCLVPEKRQEVTTEGGLDVAGQIDKMT
3F4N:A|PDBID|CHAIN|SEQUENCE         VRILRQTIQTRMNLEMAVTDEMVDIACDIKPHFCCLVPEKRQEVTTEGGLDVAGQVDKMT
                                    *****:*****************.***:*:*************************:****

tr|A1JKK6|A1JKK6_YERE8              VAVSRLAKAGILVSLFIDADMRQIDAAVAVGAPYIEIHTGAYADATSDLARQAELVRIAK
3F4N:A|PDBID|CHAIN|SEQUENCE         LAVGRLADVGILVSLFIDADFRQIDAAVAAGAPYIEIHTGAYADASTVLERQAELMRIAK
                                    :**.***..***********:********.***************:: * *****:****

tr|A1JKK6|A1JKK6_YERE8              AATYAASKGLKVNAGHGLTYHNVQPIAALPEMHELNIGHAIIGQAVMSGLAAAVTDMKVL
3F4N:A|PDBID|CHAIN|SEQUENCE         AATYAAGKGLKVNAGHGLTYHNVQPIAALPEMHELNIGHAIIGQAVMTGLAAAVTDMKVL
                                    ******.****************************************:************

tr|A1JKK6|A1JKK6_YERE8              MREARR
3F4N:A|PDBID|CHAIN|SEQUENCE         MREARR
                                    ******
```

Figure 3.11. *Alignment between template and protein pdxJ generated by clustalw reflecting a reasonable conservation.*

Table 3.5. Stereo-chemical properties of comparative homology modeled structure.

Structure Resources	Number of Residues Allowed Region	Additionally allowed region	Disallowed Region	Errat	Z-score	Verify-3D
MODELLER 1	57 (11.8%)	378 (78.4%)	47 (9.8%)	17.992	-0.29	19.34%
MODELLER 2	53 (11.0%)	388 (80.5%)	41 (8.5%)	16.318	-0.31	18.93%
MODELLER 3	65 (13.5%)	371 (77.0%	46 (9.5%)	15.063	-0.1	10.70%
MODELLER 4	69 (14.3%)	370 (76.8%)	43 (8.9%)	17.782	0.02	18.11%
MODELLER 5	61 (12.7%)	376 (78.0%)	45 (9.3%)	18.828	-0.35	30.45%
PHYRE	7 (2.9%)	230 (96.2%)	2 (0.8%)	99.142	-	91.29%
CPH-MODEL	10 (4.1%)	230 (95.4%)	1 (0.4%)	97.447	-0.1	95.88%
SWISS-MODEL	9 (1.9%)	469 (97.7%)	2 (0.4%)	99.573	-8.02	91.74%

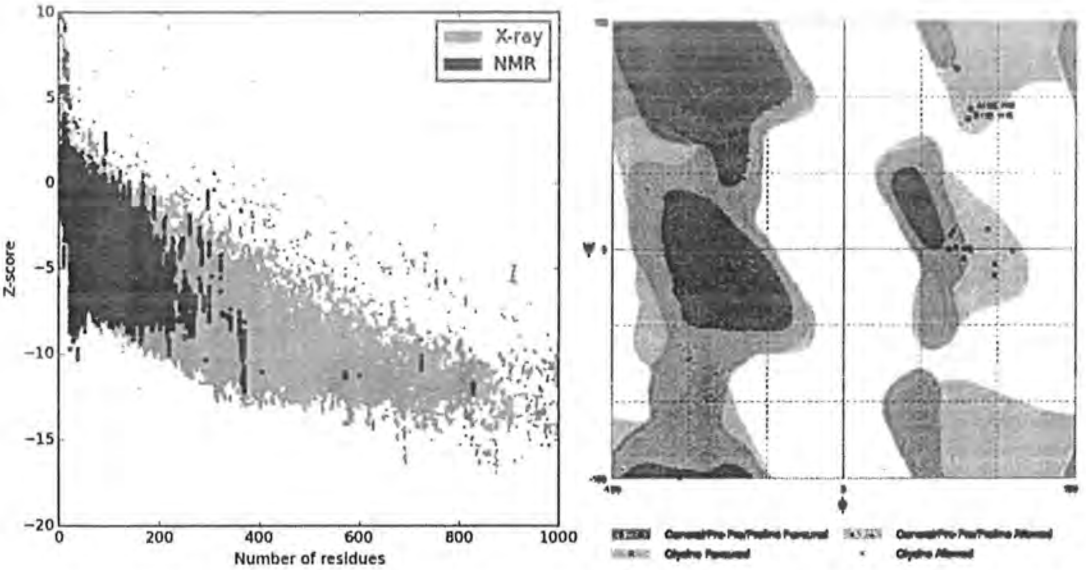


***Figure 3.12. a)** Z-score calculated from ProSa **(b)** Ramachandran plot representing Psi and Phi angles of the selected model.*

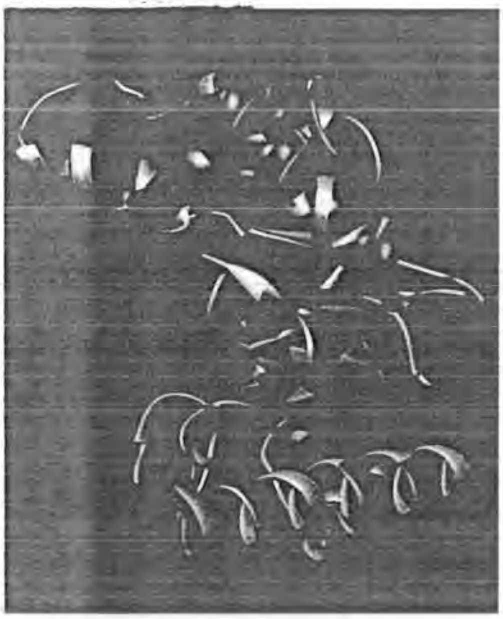


Figure 3.13. Superimposition of template and selected optimum model. Blue represents template while modeled protein is shown in green color.

3.7. Molecular Docking

The information about the domain organization within the pdxJ protein combined with the identification of the active site guided the molecular docking procedure. The steps involved in this procedure are illustrated below.

3.7.1. *Active Site identification*

Active site of pdxJ was identified via literature search. The active site of pdxJ constitutes of positive charged amino acid residue. The selected amino acid residue in the current course of study is Arg 17, illustrated in **Figure 3.14.** A conserved active site is observed in different orthologues of pdxJ.

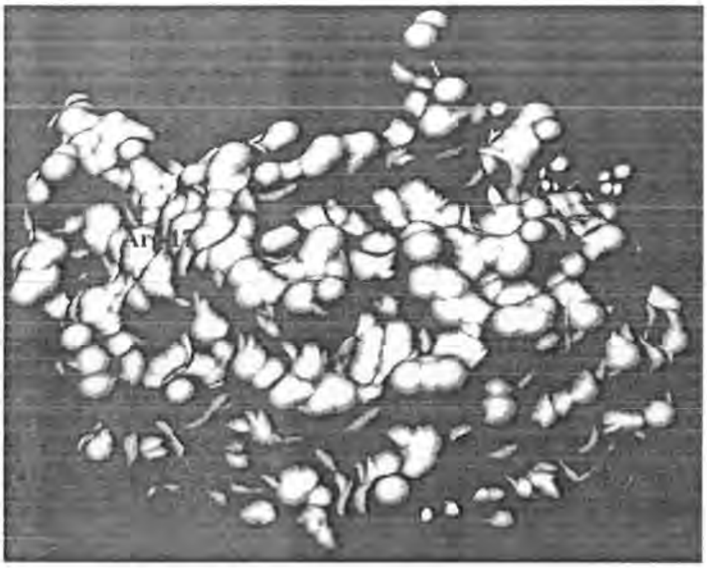


Figure 3.14. Arg17 is shown in the active site of protein.

3.7.2. *Inhibitors selection*

The inhibitors selected to be docked into the active site of pdxJ were mostly selected from different databases and their analogues were also used as potential inhibitors against pdxJ. In the current study inhibitors were accessed from pubchem. A list of 230 compounds were generated which were selected to dock into the active site of pdxJ. Natural compounds were also docked in addition to the library of synthetic compounds.

3.7.3. *Interaction Analysis*

A total of 730 ligands were docked into the active site of target, using GOLD and MOE. Natural inhibitor 2-acetyl-3-(2-heptanamidoethyl)-1H-indol-6-yl heptanoate was the top scoring compound in GOLD.

3.7.4. *Binding pattern analysis*

The prepared ligand molecules were docked into the active site of the target using GOLD. The highest Gold Score of 63 was achieved for compound 2-acetyl-3-(2-heptanamidoethyl)-1H-indol-6-yl heptanoate. Docking scores arranged in descending order of respective GOLD score values provided in **Table 3.6**. Detailed visualization analysis carried out through UCSF chimera, LIGPLOT and MOE revealed the conformational details and preferred orientations of the ligand binding. A graphical representation of the docked ligand via GOLD positioned with the active site is outlined in **Figure 3.15** and **Figure 3.16**. Hydrogen bond abetment the molecular recognition and play a pivotal role in ligand receptor interactions. LIGPLOT analysis revealed an extensive coordination system between the ligand and pdxJ. A protein residue Arg 17 forms a hydrogen bond of 2.60 A with the ligand depicted **in Figure 3.17.**

Table 3.6. Docking results of top ten docked inhibitors in descending order of GOLD Score with the corresponding binding affinities.

Sr. NO	Compound Name.	Gold Score
1	Juliflorine	70.646
2	PK000391	70.177
3	(2-decoxy-2-oxoethyl)-[2-[(2-decoxy-2-oxoethyl)-dimethylazaniumyl]ethyl]-dimethylazanium dichloride	69.412
4	(2-decoxy-2-oxoethyl)-[2-[(2-decoxy-2-oxoethyl)-dimethylazaniumyl]ethyl]-dimethylazanium	66.535
5	6'-O-(4"-Methoxy-Trans-Cinnamoyl)-Kaempferol-3-B-D-Glucopyranoside	64.868
6	Limbonin	63.923
7	2-acetyl-3-(2-heptanamidoethyl)-1H-indol-6-yl heptanoate	63.701
8	Methyl Lithospermate B	63.564
9	PK000583	61.668
10	Margosinolone	61.390

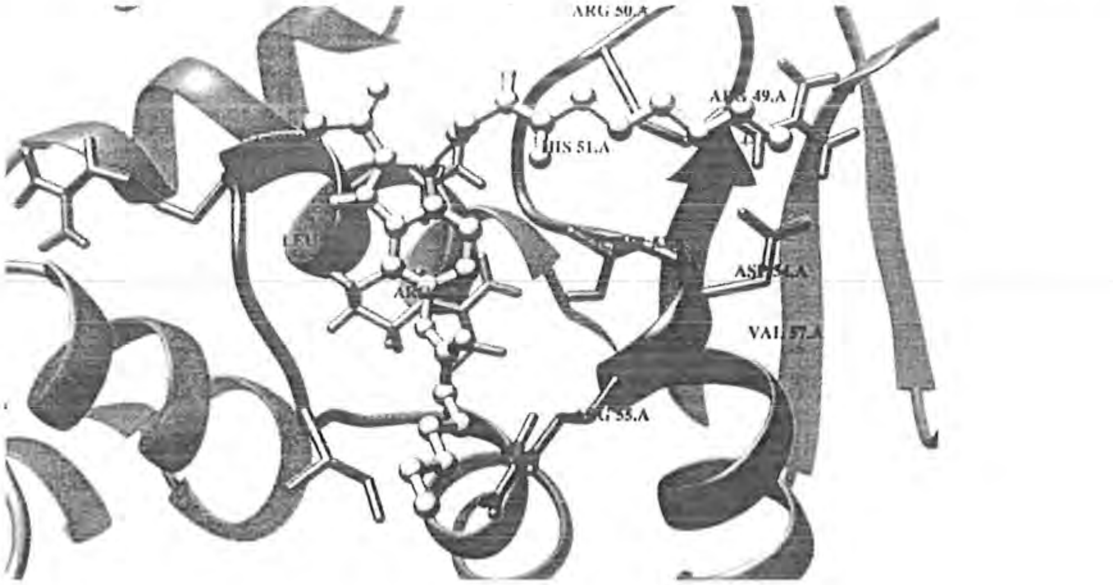


Figure 3.15. Best docked inhibitor in the active site of *pdxj protein* (GOLD).

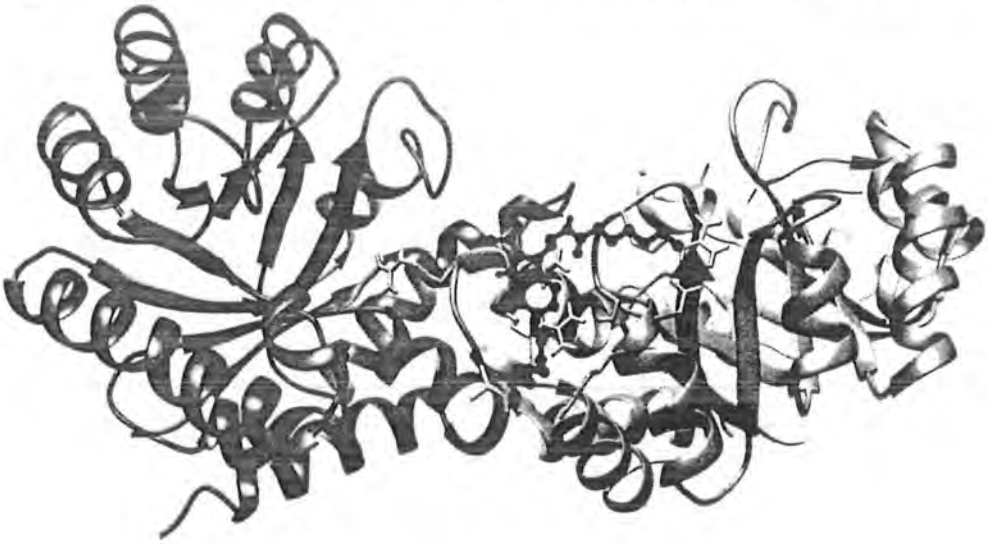


Figure 3.16. An overview of 2-acetyl-3-(2-heptanamidoethyl)-1H-indol-6-yl heptanoate docked into the active site of selected target, pdxJ (GOLD).

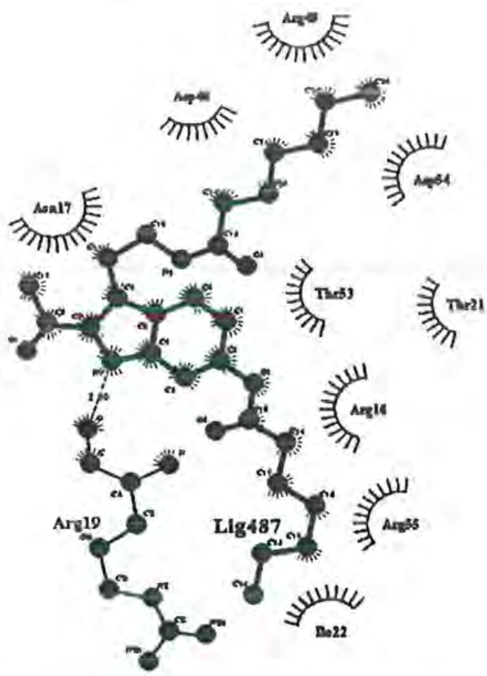


***Figure 3.17**. Interaction of ligand with pdxJ, highlighting interacting residues through LIGPLOT.*

3.8. Molecular Dynamics Simulation

The docking study provide meaningful insights into the structural basis of druggability potential of *Y.enterocolitica*. However, it provide this information within the context of static environment. In order to filtrate the dynamic behavior of protein simulation protocol was carried out followed by trajectory analysis to assess various properties of ligand bound protein. Simulation studies not only depict the dynamic behavior of protein but also highlight the important residues that play an important role in dynamic behavior. Properties including RMSD, RMSF, B-factor, and radius of gyration were plotted as a function of time to understand the biomolecular movements within a solvated environment. Analysis of protein in ligand bound form led to evaluation of structural transformation and underlying atomic level transition. It helped in revealing ligand induce variability and the dynamic role of co-factor in the presence and absence of inhibitor.

3.8.1. *Root Mean Square Deviations (RMSD)*

The deviation of the backbone Cα atoms was noticed for the entire production run of the docked protein for a time period of 100ns. The RMSD behavior of the inhibitor bound pdxj, over the studied time scale is mostly stable with an average value of 1.5Å reaching the maximum value of 2.4 at the 90th ns. The ligand placement was well complemented within the active site during the simulation and does not destabilize the protein (**Figure 3.18**).

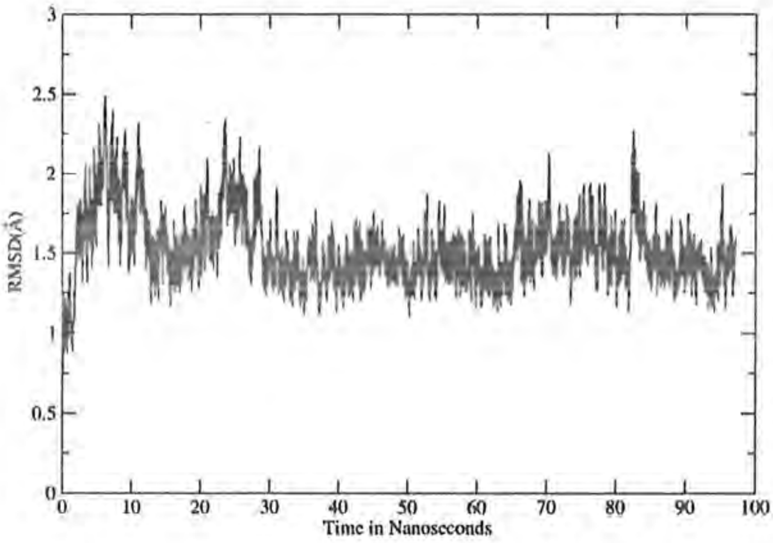


Figure 3.18. *Root Mean Square Deviation RMSD plots of docked protein complex for 100 ns simulation run.*

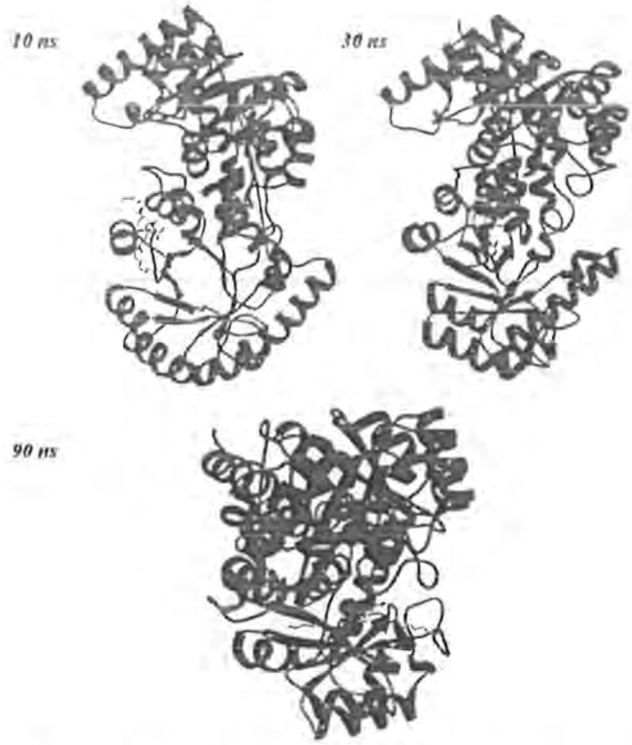


Figure 3.19. *Snapshots of docked pdxj over a time lapse of 10ns, 30ns, and 90ns. The helices are depicted in purple color, sheets in red and the loops in sea blue color.*

3.8.2. *Root Mean Square Fluctuations (RMSF)*

As a measure of atomic fluctuation, RMSF provides a mean to recognize and comprehend the structurally flexible and rigid regions of the drug target. The average Cα fluctuation for the ligand bound protein was observed to be 0.9 Å. The maximum value of RMSF for the ligand bound protein was 3.2Å. Detailed analysis of trajectories lead to identification of those protein substructures that are responsible for obtained RMSF trend. The most important observation in this context was the higher fluctuations observed for region that are involve in loop and turns forming and are solvent exposed. The active site residue found to have an average RMSF value less than 2 Å signifying the stability during the production run (**Figure 3.20**).

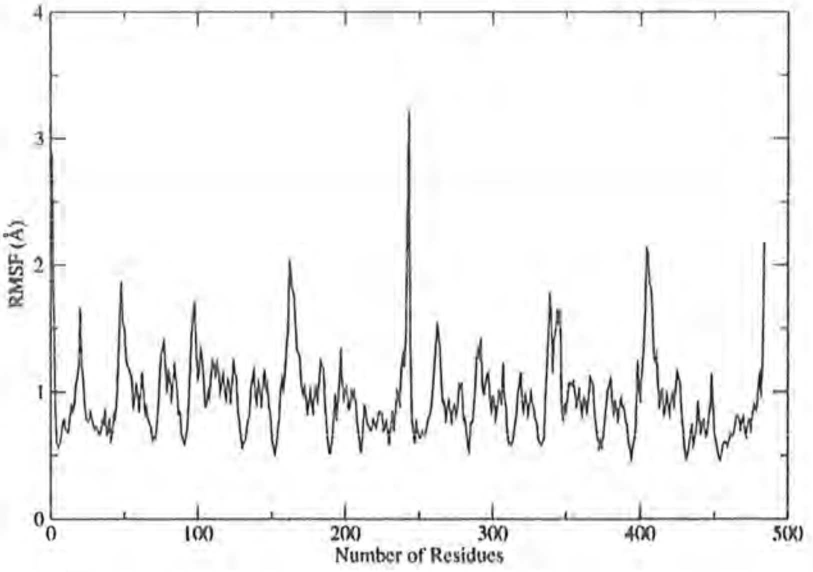


Figure 3.20. Root Mean Square fluctuation RMSF of docked protein-ligand complex over 100 ns simulation run

3.8.3. *β-Factor Analysis*

β-Factor explains the flexibility and thermal stability of protein over some time period. β-Factor is a quantity that is measured in terms of RMSF. Its value is therefore dependent on the level of localized atomic fluctuations which collectively contribute to the global vibrational movements of the protein and its thermal stability. The pattern of B-factor for protein is consistent with the RMSF trend. β-Factor average value calculated for docked complex was 27Å and plot demonstrated the higher instability is seen on 243th residues of protein (**Figure 3.21**). Other fluctuations observed on the residue no 0.3, 161, 404, and 483 having value less than 150Å.

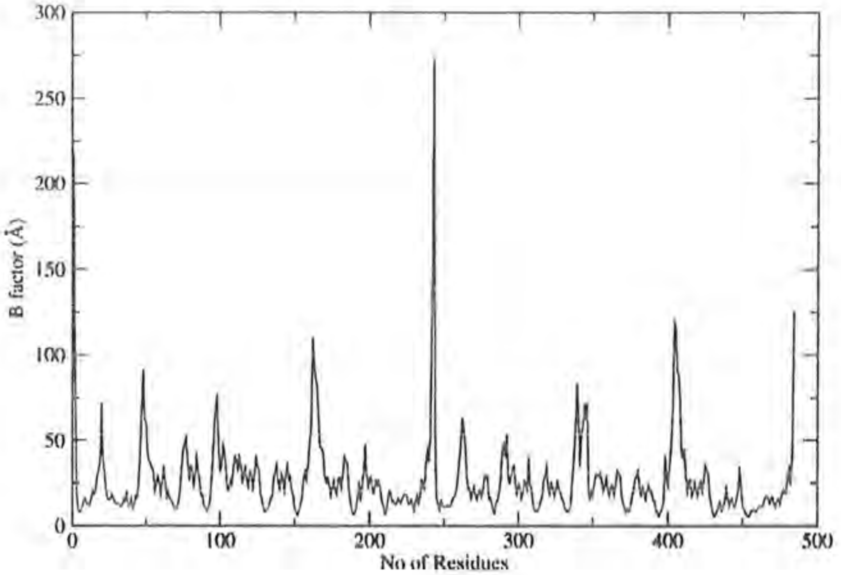


Figure 3.21. β-Factor graph of protein ligand complex over 100ns simulation run.

3.8.4. *Radius of Gyration (Rg)*

To evaluate the structural compactness, radius of gyration was calculated as a time function for the 100 ns simulation of ligand-protein complex. The average value of 25.3 Å for the docked protein (Figure 16) denotes the stability of the protein structure (**Figure 3.22**).

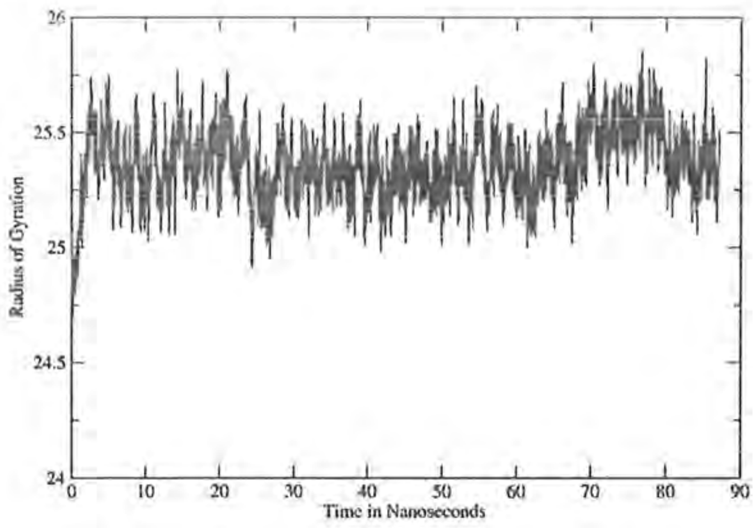
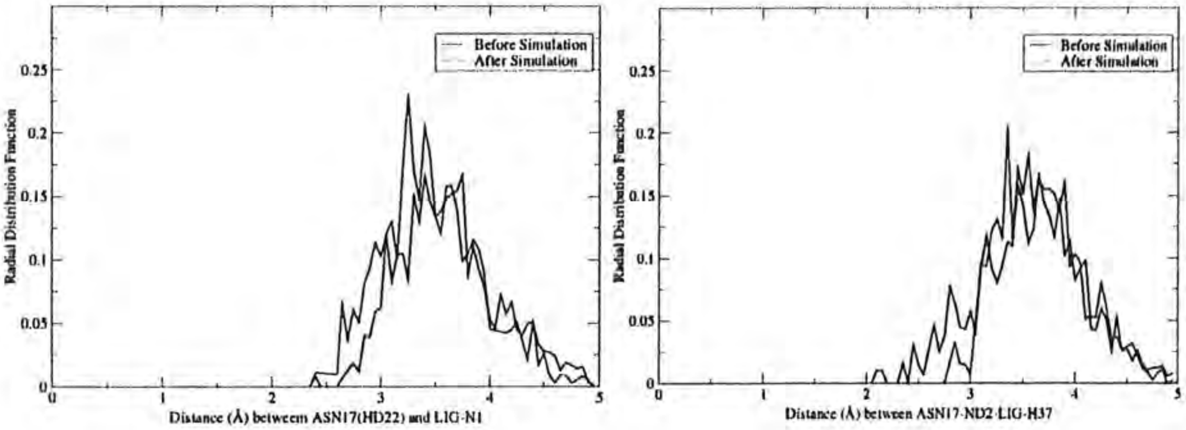


Figure 3.22. Radius of gyration of protein ligand docked complex over 100 ns simulation time period.

3.9. Radial Distribution Function Analysis

RDF graphs depict the location of residues and molecules present either in the interior or exterior of the system. It helps in finding out the distribution of different type's atoms, molecules, and species of the receptor protein residues which play crucial roles in ligand binding and stability. For this, hydrogen bonding calculation was done to find hydrogen interactions that tightly bound the inhibitor over the course of the simulation period. The RDF between the HD22, ND2, OD1 of ASN17 and the ligand was calculated during the course of 100 ns simulation to understand to dynamics of this catalytic residues. Similarly, RDF graph was also generated for the ASN17 hydrogen atom and the ligand for understanding its dynamics. In the case of the HD22 atom of ASN17 and ligand H37, the highest peak was observed at 2.7 Å with g (r) value of 0.27 while after simulation the highest peak was observed at 3.8 Å with g (r) value of 0.15. The figure suggests that the graph is more refine and tend to narrow down after simulation, indicating that the distance between Asn17 HD22 atom and ligand H37 decreases. It further means that the strength hydrogen interaction between these two residues increases after simulation period (**Figure 3.23A**). Respectively in case of ASN17 HD22 atom and ligand N atom. Before simulation, highest peak was observed at 3.7 Å with g (r) value of 0.1 and average distance of 2.6 while after simulation the highest peak was observed at 3.2 Å with g (r) value of 0.2 and average distance of 2.6 (**Figure 3.23B**). For ND2 atom of Asn17 and the ligand, before simulation, the highest peak was observed at 3.6 Å with g (r) value of 0.16 while after simulation the highest peak was found at 3.3 Å with g (r) value of 0.20. The peak before simulation was greater in magnitude than after simulation and the average distance after simulation was smaller showing that the strength hydrogen interaction between these two residues increases after simulation period (**Figure 3.23C**). In case of ASN OD1 atom and ligand N atom before simulation, highest peak was observed at 2.9 Å with g (r) value of 0.05 and average distance of 2.3 while after simulation the highest peak was observed at 3.6 Å with g (r) value of 0.1 and average distance of 2.3 (**Figure 3.23D**). For OD1 atom of Asn and the ligand, before simulation, the highest peak was observed at 3.7 Å with g (r) value of 0.1 while after simulation the highest peak was found at 3.7 Å with g (r) value of 0.1. Whereas the observed distance between OD1 atom of Arg and ligand was less after simulation showing the increase in strength of interaction between two residues after simulation thus maintaining the overall structure of complex (**Figure 3.23E**).



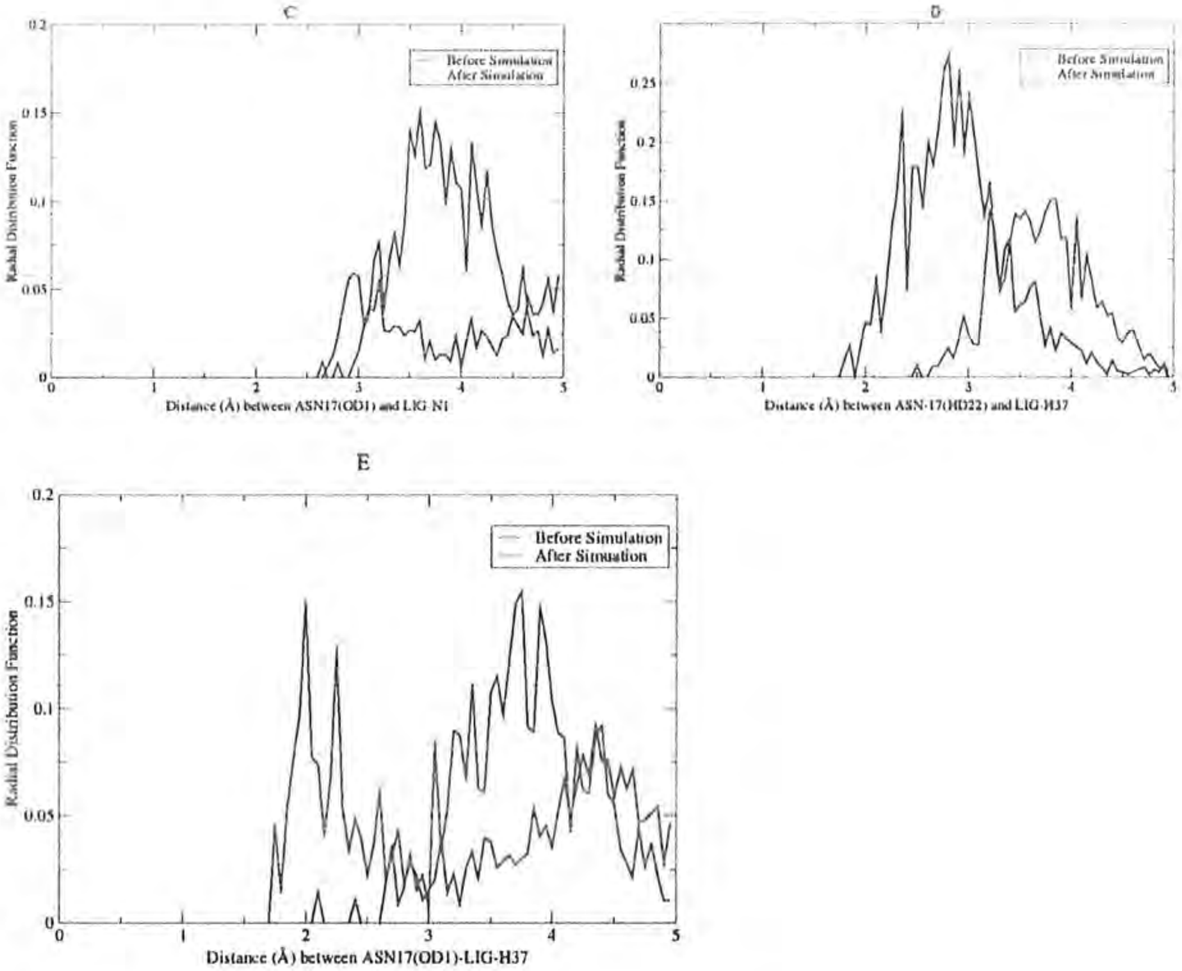


Figure 23. A. RDF graph of Asn (HD22) and ligand (H37).**B.** RDF graph of Asn (HD22) and ligand (nitrogen).**C.** RDF graph of Asn17 (ND2) and ligand (H37). **D.** RDF graph of Asn (OD1) and ligand (nitrogen). E. RDF graph of Asn (OD1) and ligand (Hydrogen).

Discussion:

Infectious diseases are the leading and eminent threat for the public health worldwide. Ability of adaptation and reshaping of pathogenic mechanism in bacterial species limits the activity of antimicrobial agents against these pathogens. The emergence of antibiotic resistance strains are not just because of high rate of bacterial genome reshaping by mutations but also triggered by antibiotics imposed selective pressures. Thus, all these bottlenecks demand newer, specific and more therapeutic agents. Progressive expansion of bioinformatics based *in silico* techniques successfully assist in overcoming different obstacles in drug and vaccine discovery process. The current study emphasis on the identification of prospective vaccine and druggable candidates against *Y.enterocolitica* based on proteome-wide screening approach.

MeoA and OmpC2 proteins were found to be an effective vaccine candidates against *Y. enterocolitica*. Both proteins are declared as human non-homologs thus avoid the chances of autoimmune responses and cross reactivity between the host and pathogen proteins (Barrios *et al.,* 1994). They are found to be essential targets in inhibition of pathogen growth and limit the survival chances in competing environment of the host. Virulent protein analysis demonstrated that both proteins could evoke strong signaling pathways in the host (Cornelis *et al.*, 1989). Proteins localized in the outer membrane or extracellular matrix can act as possible therapeutic targets as such protein come in frequent contact to biotic and abiotic factors of the extracellular environment and mediating immune responses. Targeting such proteins are also significant from their role in pathogen adherence, invasion and proliferation in host cells (Forman *et al.*, 2008). Furthermore, the proteins found to have low molecular weight and possess the least number of transmembrane helices. These parameters prioritized both the candidates as ideal proteins for experimental analysis due to easy purification, cloning and expression analysis. The proteins were further found as an adhesive, which is an important factor as adhesive proteins aid in bacterial adherence to host cells and subsequent colonization and infection (Maritz & Richards 2014).

The meoA protein is mapped for 'YGADNFLSQ' epitope. This epitope is highly antigenic (1.06) with the potential of evoking both humoral and cell-mediated immunity. The peptide has shown high affinity toward MHC alleles (Total alleles binding = 16) and low IC50 value for the DRB1*0101 allele (36 nM). In addition, the predicted epitope was found non-allergen and completely conserved among nine strains of the pathogen. Likewise, OmpC2 has mapped for "YASSNRTTA" epitope. The highly antigenic nature (1.06) of the epitope with high affinity for MHC alleles (Total alleles binding =16) and low IC50 value for the DRB1*0101 allele (22 nM) make it an excellent target. Epitopes are further declared as non-allergen and highly conserved, suggesting it a potential target for designing an effective vaccine. Both the proteins were assessed for cellular interactome to decipher their interacting network. It was revealed that both the proteins interact with many vital proteins of *Y. enterocolitica*, indispensable for the pathogen survival and host. The three-dimensional structure of proteins was predicted in order to view the topology of the screened epitopes. The epitopes of proteins disclose exomembrane topology and were not folded in the protein globular structure. Molecular docking analysis of the epitope showed epitopes possess high binding affinity towards the binding pocket of DRB1*0101 allele and formed multiple hydrogen bonds suggesting formation of stable complexes. The phenomenon of hydrogen bond formation is significant in folding and stability. MeoA has not yet used as vaccine target while OmpC2 has been used as a multivalent vaccine composition

The *in silico* based approach involves a series of screening steps to find a potential drug targets (Butt *et al.*, 2012). The identification of pdxj protein as novel drug targets in the present study provides a basis for computer-aided drug design against *Y.enterocolitica* to overcome the challenges of severe infections. For drug target identification an *in silico* subtractive genomic approach was applied to screen pathogen reference genome 8081, contain total of 3957 proteins followed by the application of subtractive genomic steps. Duplication sequences from pathogenic genome are removed by the tool known as CD-HIT to facilitate the binding of drug molecule. During pathway analysis total 414 unique bacterial pathways were identified which could be targeted for therapeutic effect. Among the unique bacterial pathways higher numbers of genes were present in two-component system, Biosynthesis of secondary metabolites, amino acid and nucleotide sugar metabolism, starch and sucrose metabolism, fructose and mannose metabolism and microbial metabolism in diverse environment. Further strict filters are applied to ensure the existence of only non-homologous sequences of pathogen are filtered out by BLASTp against RefSeq database. Homology search is carried out to explore the essentiality of non-homologous sequences for bacteria. Pathogen's essential protein has an impact on the survival of pathogen as their inhibition within host system can put a halt on bacterial growth. The identified unique proteins are then subjected to two fold round of analysis where assessment is done on the basis of druggability followed by subcellular localization analysis revealed 121 cytoplasmic bacterial enzymatic proteins. Proper localization of protein in living cell directly influences the protein functionality. Target localization highly influences the binding of a drug to its target as cytoplasmic proteins can act as possible therapeutic targets (Barh *et al.* 2009). Based on these factors, after evaluation of catalytic and structural requirements pdxj protein was selected as therapeutic candidate to be subjected to modelling procedure. Pdxj is the active form of vitamin B6 that acts as an essential, ubiquitous coenzyme in amino acid metabolism.

Homology modelling was performed to model structure and evaluated with different online tools which tested the quality of structures generated. Model generated by Swiss-model was selected as best model for conducting further analysis as it had minimal bad contacts. Pairwise sequence alignment of pdxj with their respective template yielded a highly conserved pattern of similarity. This likeness concur the correct selection of template, similar regions and evolutionary conservation of sequences. Generated protein model found to have high propensity of alpha helices and less number of beta sheets. The monomer of PNP synthase consists of one compact domain that adopts the abundant TIM barrel fold which is consider as one of the most common conserved protein fold. It control the enzymatic catalysis and maintain the structure of protein.

The homology modelling served as a starting point for the docking and subsequent simulation procedure. The ligand molecule docked into the active site of the target using GOLD and MOE. Total of 500 natural and 230 synthetic compounds were docked into the active site of protein, using GOLD and MOE. The highest GoldScore of 63.7 was exhibited by natural inhibitor 2-acetyl-3-(2-heptanamidoethyl)-1H-indol-6-yl heptanoate. In addition the preferred ligand binding modes highlight the chemical basis of interaction and type of inhibition. The phenomenon of hydrogen bond formation has an immense importance in structure stability (Saenger & Jeffrey 1991). Within the active site ligand was found to have hydrogen bond with an active site residue Arg 19 with 2.60 Å distance. The docking study further needed an understanding of the structural adjustment made upon ligand introduction into the system. The next phase of molecular dynamic simulation not only facilitates the drug design approach but also reflect the time dependent behavior of system (Suresh *et al.*, 2008). Water hold significant biological importance in chemical processes especially in the context of protein dynamics. Activity and dynamic behavior of protein is greatly influences by the presence of water thus, dominating the catalytic activity and functional inhibition. The application of simulation to the *Y.enterocolitica* thus explains the system stability over the studied time scale.

The structural parameters focused in our current study represent the stability of complex. Stability of docked protein complex was explained by calculation of physical properties over a time period of 100ns by keeping four parameters in view i.e. RMSD, RMSF, Radius of gyration, B-factor. The RMSD behavior of the inhibitor bound pdxj, over the studied time scale was mostly stable with an average value of 3.2 Å with no major peaks, reflecting the consistent maintenance of overall protein architecture. The average radius of gyration also validates this observation. *In silico* line of work adopted during the current project provided meaningful information at various stages of analysis. Average RMSF values was 3.2 with maximum fluctuation up to 3.0 Å to 3.5 Å in loop region. As they are allowed to move freely they tend to fluctuate more than the rest of residues. Structural changes observed in docked complex within the course of 100 ns simulation includes extension of α-helices and β-sheets. Within the binding site of the protein, even under the dynamics conditions ligand has remained strongly bound within the binding pocket of pdxj validating its role as a competitive inhibitor. Small scale reorientation of the inhibitor molecule has taken place which made the inhibitor molecule to become closer in the vicinity of protein active site with an increase in number of hydrogen bond interaction thus making the overall complex more stable.

In silico strategy adopted during the current project provided eloquent information at various stages of analysis. Collectively, the inferred knowledge about essential catalytic mechanism and effects of chemical and structure variability on inhibitor binding can be extended to increase efficacy and potency of novel *Y.enterocolitica* 2-acetyl-3-(2-heptanamidoethyl)-1H-indol-6-yl heptanoate natural inhibitor in order to halt the lethal infection caused by the studies pathogen.

Conclusion

In silico subtractive genomics is a rapid, powerful, and cost-effective approach for screening of drug and vaccine targets for any given pathogen, provided both the pathogen and host genomes are available. In this integrated analysis, we concurred the identification of OmpC2 and meoA proteins as novel epitope-based vaccine candidates and pdxj as potential drug candidates against *Y. enterocolitica*.

In reverse vaccinology protocol both the proteins are pathogen-specific and crucial for its survival in competing host environment. Subcellular localization revealed their exomembrane topology, an important consideration for evoking and interacting with host immune system products. The proteins are recognized as virulent and could mediate strong signaling pathways in the host. They are found to be ideal candidates for investigating their immune protection efficacies in animal models due to their suitable physicochemical properties. PPIs decipher their strong interactions with proteins involved in major metabolic pathways of the pathogen. Furthermore, the proteins are antigenic and mapped for conserved, virulent and surface-exposed antigenic 9mer epitopes which docked deeply in binding cavity of the DRB*0101 allele, most prevalent allele in humans. Collectively, the findings of this study could provide the foundation for designing peptide or recombinant vaccine against *Y. enterocolitica*.

Exploration of lead compound PK000697 against pdxj protein can be used as potential drug target compound against this pathogenic bacteria in order to block the infection it causes. Molecular docking protocol classified 2-acetyl-3-(2-heptanamidoethyl)-1H-indol-6-yl heptanoate *as* the putative druggable compound. An essential non-homologous protein that has no counterpart in human. Comparative homology modelling was applied to attain a high quality model for structurally uncharacterized pdxj. MD simulations lead to the findings that the protein undergoes some conformational changes explaining its dynamic behavior with respect to time. Besides the side chain fluctuations and a sheet to loop transformation, stability of inhibitor and target protein complex was observed. The stability and interaction between pdxj and best docked ligand in solvated system is concurred by the molecular docking simulation procedure. Analysis involving RDF illustrated that ASN17-HD22, ASN17-ND2, ASN17-OD1, ASN-HD22 and ASN-OD1 were the main residue involved in stable interaction with LIG-N, LIG-H, LIG-N, LIG-H, and LIG-H. It can be concluded that molecular dynamics can enhance the existing pharmacological designs to develop more potent, specific and efficient drugs against MDR *Y. enterocolitica*.

REFERENCES

Abadio, A. K. R., Kioshima, E. S., Teixeira, M. M., Martins, N. F., Maigret, B., & Felipe, M. S. S. (2011). Comparative genomics allowed the identification of drug targets against human fungal pathogens. *BMC genomics, 12*(1), 75.

Alonso, H., Bliznyuk, A. A., & Gready, J. E. (2006). Combining docking and molecular dynamic simulations in drug design. *Medicinal Research Reviews, 26*(5), 531-568.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410.

Amineni, U., Pradhan, D., & Marisetty, H. (2010). *In silico* identification of common putative drug targets in Leptospira interrogans. *Journal of chemical biology, 3*(4), 165-173.

Ariel, N., Zvi, A., Grosfeld, H., Gat, O., Inbar, Y., Velan, B., & Shafferman, A. (2002). Search for potential vaccine candidate open reading frames in the Bacillus anthracis virulence plasmid pXO1: in silico and in vitro screening. *Infection and Immunity, 70*(12), 6817-6827.

Azam, S. S., Uddin, R., & Wadood, A. (2012). Structure and dynamics of alpha-glucosidase through molecular dynamics simulation studies. *Journal of Molecular Liquids, 174*, 58-62.

Barh, D., & Kumar, A. (2009). In silico identification of candidate drug and vaccine targets from various pathways in Neisseria gonorrhoeae. *In silico biology, 9*(4), 225-231.

Barh, D., Misra, A. N., Kumar, A., & Vasco, A. (2010). A novel strategy of epitope design in Neisseria gonorrhoeae. *Bioinformation, 5*(2), 77.

Barh, D., Misra, A. N., Kumar, A., & Vasco, A. (2010). A novel strategy of epitope design in Neisseria gonorrhoeae. *Bioinformation*, *5*(2), 77.

Barh, D., Tiwari, S., Jain, N., Ali, A., Santos, A. R., Misra, A. N., Kumar, A. (2011). In silico subtractive genomics for target identification in human bacterial pathogens. *Drug Development Research*, *72*(2), 162-177.

Barrios, C. (1994). Heat shock proteins as carrier molecules: in vivo helper effect mediated by Escherichia coli GroEL and DnaK proteins requires cross-linking with antigen. *Clin. Exp. Immunol.*, *98*, 175-177.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., & Tasumi, M. (1977). The protein data bank. *The FEBS Journal*, *80*(2), 319-324.

Betts, J. C. (2002). Transcriptomics and proteomics: tools for the identification of novel drug targets and vaccine candidates for tuberculosis. *IUBMB life*, *53*(4-5), 239-242.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., & Schwede, T. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, *42*(W1), W252-W258.

Bottone, E. J. (1997). Yersinia enterocolitica: The charisma continues. *Clinical Microbiology Reviews*, *10*(2), 257-276.

Bottone, E. J. (1997). Yersinia enterocolitica: The charisma continues. *Clinical Microbiology Reviews*, *10*(2), 257-276.

Bottone, E. J., & Mollaret, H. H. (1977). Yersinia enterocolitica: a panoramic view of a charismatic microorganism. *Critical Reviews in Microbiology*, *5*(2), 211-241.

Butler, T., Islam, M., Islam, M. R., Azad, A. K., Huq, M. I., Speelman, P., & Roy, S. K. (1984). Isolation of Yersinia enterocolitica and Y. intermedia from fatal cases of diarrhoeal illness in Bangladesh. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *78*(4), 449-450.

Butt, A. M., Tahir, S., Nasrullah, I., Idrees, M., Lu, J., & Tong, Y. (2012). Mycoplasma genitalium: a comparative genomics study of metabolic pathways for the identification of drug and vaccine targets. *Infection, Genetics and Evolution*, *12*(1), 53-62.

Butt, A. M., Tahir, S., Nasrullah, I., Idrees, M., Lu, J., & Tong, Y. (2012). Mycoplasma genitalium: a comparative genomics study of metabolic pathways for the identification of drug and vaccine targets. *Infection, Genetics and Evolution*, *12*(1), 53-62.

Cavasotto, C. N., & Phatak, S. S. (2009). Homology modeling in drug discovery: current trends and applications. *Drug Discovery Today*, *14*(13), 676-683.

Chen, L., Zheng, D., Liu, B., Yang, J., & Jin, Q. (2014). VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. *Nucleic Acids Research*, *44*(D1), D694-D697.

Chen, X., Ji, Z. L., & Chen, Y. Z. (2006). TTD: therapeutic target database. *Nucleic Acids Research*, *30*(1), 412-415.

Chen, X., Ji, Z. L., & Chen, Y. Z. (2016). TTD: therapeutic target database. *Nucleic Acids Research*, *30*(1), 412-415.

Chou, P. Y. (1989). Prediction of protein structural classes from amino acid compositions. In *Prediction of protein Structure and the Principles of Protein Conformation* (pp. 549-586).

Colovos, C., & Yeates, T. O. (1993). ERRAT: an empirical atom-based method for validating protein structures. *Protein Sci*, *2*, 1511-1519.

Cornelis, G. R., Biot, T., Rouvroit, C. L., Michiels, T., Mulder, B., Sluiters, C., & Vanooteghem, J. C. (1989). The Yersinia yop regulon. *Molecular Microbiology*, *3*(10), 1455-1459.

Cover, T. L., & Aber, R. C. (1989). Yersinia enterocolitica. *New England Journal of Medicine*, *321*(1), 16-24.

Dimitrov, I., Bangov, I., Flower, D. R., & Doytchinova, I. (2014). AllerTOP v. 2—a server for in silico prediction of allergens. *Journal of Molecular Modeling*, *20*(6), 2278.

Doytchinova, I. A., & Flower, D. R. (2007). VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, *8*(1), 4.

El-Maraghi, N. R., & Mair, N. S. (1979). The histopathology of enteric infection with Yersinia pseudotuberculosis. *American journal of clinical pathology*, *71*(6), 631-639.

Emini, E. A., Hughes, J. V., Perlow, D., & Boger, J. (1985). Induction of hepatitis a virus-neutralizing antibody by a virus-specific synthetic peptide. *Journal of Virology*, *55*(3), 836-839.

Forman, S., Wulff, C. R., Myers-Morales, T., Cowan, C., Perry, R. D., & Straley, S. C. (2008). yadBC of Yersinia pestis, a new virulence determinant for bubonic plague. *Infection and Immunity*, *76*(2), 578-587.

Fredriksson M., Linström, M., and Korkeala, H. (2009). "Yersinia enterocolitica and Yersinia Pseudotuberculosis," in pathogens and toxins in foods: challenges and interventions. *J. Sofos*.

Garg, A., & Gupta, D. (2008). VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics*, *9*(1), 62.

Garg, A., & Gupta, D. (2008). VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics*, *9*(1), 62.

Georrge, J. J., & Umrania, V. (2011). In silico identification of putative drug targets in Klebsiella pneumonia MGH78578.

Giuliani, M. M., Adu-Bobie, J., Comanducci, M., Aricò, B., Savino, S., Santini, L., & Cartocci, E. (2006). A universal vaccine for serogroup B meningococcus. *Proceedings of the National Academy of Sciences*, *103*(29), 10834-10839.

Goodfellow, J. M. (1990). Molecular dynamics: Application in Molecular Biology. In: Good fellow, J. M. ed. London: CRC Press.

Goldman, M., & Blajchman, M. A. (1991). Blood product-associated bacterial sepsis. *Transfusion medicine reviews*, *5*(1), 73-83.

Griffin P. M., Carneil E. (2014). Control of Communicable Diseases Manual, 20th ed. Washington, D.C.: American Public Health Association. *Yersiniosis*, 690–3.

Hansson, T., Oostenbrink, C., & van Gunsteren, W. (2002). Molecular dynamics simulations. *Current Opinion in Structural Biology*, *12*(2), 190-196.

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., & Steinbeck, C. (2015). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, *44*(D1), D1214-D1219.

Hermann, E., Lohse, A. W., Mayet, W. J., ZEE, R., EDEN, W., Probst, P., & Fleischer, B. (1992). Stimulation of synovial fluid mononuclear cells with the human 65-kD heat shock protein or with live enterobacteria leads to preferential expansion of TCR-γδ+ lymphocytes. *Clinical & Experimental Immunology*, *89*(3), 427-433.

Hong, L., & Lei, J. (2009). Scaling law for the radius of gyration of proteins and its dependence on hydrophobicity. *Journal of Polymer Science Part B: Polymer Physics*, *47*(2), 207-214.

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, *26*(5), 680-682.

Huang, Y., Niu, B., GAO, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, *26*(5), 680-682.

Jones, D. B., Coulson, A. F., & Duff, G. W. (1993). Sequence homologies between hsp60 and autoantigens. *Immunology today*, *14*(3), 115-118.

Kadam, K., Sawant, S., Jayaraman, V. K., & Kulkarni-Kale, U. (2016). Databases and Algorithms in Allergen Informatics. In *Bioinformatics-Updated Features and Applications*. InTech.

Kuzmanic, A., & Zagrovic, B. (2010). Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. *Biophysical journal*, *98*(5), 861-871.

Kuzmanic, A., & Zagrovic, B. (2010). Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. *Biophysical journal*, *98*(5), 861-871.

Laskowski, R. A., & Swindells, M. B. (2011). LigPlot+: multiple ligand–protein interaction diagrams for drug discovery.

Lee, H., Heo, L., Lee, M. S., & Seok, C. (2015). GalaxyPepDock: a protein–peptide docking tool based on interaction similarity and energy optimization. *Nucleic Acids Research*, *43*(W1), W431-W435.

Lee, L. A., Gerber, A. R., Lonsway, D. R., Smith, J. D., Carter, G. P., Puhr, N. D., & Tauxe, R. V. (1990). Yersinia enterocolitica O: 3 infections in infants and children, associated with the household preparation of chitterlings. *New England Journal of Medicine*, *322*(14), 984-987.

Luthy, R., Bowie, J. U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, *356*(6364), 83.

Manque, P. A., Tenjo, F., Woehlbier, U., Lara, A. M., Serrano, M. G., Xu, P., & Buck, G. A. (2011). Identification and immunological characterization of three potential vaccinogens against Cryptosporidium species. *Clinical and Vaccine Immunology*, *18*(11), 1796-1802.

Maritz, C., & Richards, S. (2014). Considerations for Vaccine Design in the Postgenomic Era. In *Molecular Vaccines* (pp. 677-696). Springer International Publishing.

Maritz, C., & Richards, S. (2014). Considerations for Vaccine Design in the Postgenomic Era. In *Molecular Vaccines* (pp. 677-696). Springer International Publishing.

Martínez, P. O., Fredriksson-Ahomaa, M., Sokolova, Y., Roasto, M., Berzins, A., & Korkeala, H. (2009). Prevalence of enteropathogenic Yersinia in Estonian, Latvian, and Russian (Leningrad region) pigs. *Foodborne pathogens and disease*, *6*(6), 719-724.

Masignani, V., Neves, R.P.P., Martins, S.A. (2007) Yersinia. *EFSA Journal*, 130, 190–195.

Mayrose, I., Penn, O., Erez, E., Rubinstein, N. D., Shlomi, T., Freund, N. T., & Martz, E. (2007). Pepitope: epitope mapping from affinity-selected peptides. *Bioinformatics*, *23*(23), 3244-3246.

Menzies, B. E. (2010). Axillary Abscess due to Yersinia enterocolitica. *Journal of Clinical Microbiology*, *48*(9), 3438-3439.

Miller, V. L., Farmer, J. J., Hill, W. E., & Falkow, S. (1989). The ail locus is found uniquely in Yersinia enterocolitica serotypes commonly associated with disease. *Infection and Immunity*, *57*(1), 121-131.

Miller, V. L., Farmer, J. J., Hill, W. E., & Falkow, S. (1989). The Ail Locus is found uniquely in Yersinia enterocolitica Serotypes commonly associated with Disease. *Infection and Immunity*, *57*(1), 121-131.

Mora, M., Donati, C., Medini, D., Covacci, A., & Rappuoli, R. (2006). Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach. *Current Opinion in Microbiology*, *9*(5), 532-536.

Morris, G. M., & Lim-Wilby, M. (2008). Molecular docking. *Molecular Modeling of Proteins*, 365-382.

Naz, A., Awan, F. M., Obaid, A., Muhammad, S. A., Paracha, R. Z., Ahmad, J., & Ali, A. (2015). Identification of putative vaccine candidates against Helicobacter pylori exploiting exoproteome and secretome: a reverse vaccinology based approach. *Infection, Genetics and Evolution*, *32*, 280-291.

Noll, A., Roggenkamp, A., Heesemann, J., & Autenrieth, I. B. (1994). Protective role for heat shock protein-reactive alpha beta T cells in murine yersiniosis. *Infection and immunity*, *62*(7), 2784-2791.

Ochman, H., Soncini, F. C., Solomon, F., & Groisman, E. A. (1996). Identification of a pathogenicity island required for Salmonella survival in host cells. *Proceedings of the National Academy of Sciences*, *93*(15), 7800-7804.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *27*(1), 29-34.

Okwori, A. E., Martínez, P. O., Fredriksson, M., Agina, S. E., & Korkeala, H. (2009). Pathogenic Yersinia enterocolitica 2/O: 9 and Yersinia pseudotuberculosis 1/O: 1 strains isolated from human and non-human sources in the Plateau state of Nigeria. *Food Microbiology*, *26*(8), 872-875.

Parker, J. M. R., Guo, D., & Hodges, R. S. (1986). New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry*, *25*(19), 5425-5432.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605-1612.

Pieper, U., Eswar, N., Webb, B. M., Eramian, D., Kelly, L., Barkan, D. T., & Davis, F. P. (2008). MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic acids research*, *37*(suppl_1), D347-D354.

Pizza, M., Scarlato, V., Masignani, V., Giuliani, M. M., Arico, B., Comanducci, M., & Galeotti, C. L. (2000). Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science*, *287*(5459), 1816-1820.

Pockley, A. G. (2002). Heat shock proteins, inflammation, and cardiovascular disease. *Circulation*, *105*(8), 1012-1017.

Rahman, A., Bonny, T. S., Stonsaovapak, S., & Ananchaipattana, C. (2011). Yersinia enterocolitica: Epidemiological studies and outbreaks. *Journal of pathogens*, *2011*.

Ramachandran, G. N., Venkatachalam, C. M., & Krimm, S. (1966). Stereochemical Criteria for Polypeptide and Protein Chain Conformations: III. Helical and Hydrogen-Bonded Polypeptide Chains. *Biophysical journal*, *6*(6), 849-872.

Rappuoli, R. (2000). Reverse Vaccinology. *Current Opinion in Microbiology*, *3*(5), 445-450.

Rueckert, C., & Guzmán, C. A. (2012). Vaccines: from empirical development to rational design. *PLoS Pathogens*, *8*(11), e1003001.

Rueckert, C., & Guzmán, C. A. (2012). Vaccines: from empirical development to rational design. *PLoS Pathogens*, *8*(11), e1003001.

Saenger, W. and Jeffrey, G.A. (1991). Hydrogen Bonding in Biological Structures. Springer-Ver-lag, Berlin.

Sachdeva, G., Kumar, K., Jain, P., & Ramachandran, S. (2004). SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics*, *21*(4), 483-491.

Serruto, D., & Rappuoli, R. (2006). Post-genomic vaccine development. *FEBS Letters*, *580*(12), 2985-2992.

Singh, H., & Raghava, G. P. S. (2003). ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics*, *19*(8), 1009-1014.

Soltan Dallal, M. M., & Moezardalan, K. (2004). Frequency of Yersinia species infection in Paediatric acute diarrhoea in Tehran.

Sonnhammer, E. L., Von Heijne, G., & Krogh, A. (1998, July). A hidden Markov model for predicting transmembrane helices in protein sequences. In *Ismb* (Vol. 6, pp. 175-182).

Sousa, S.F., Ribeiro, A.J.M., Coimbra, J.T.S., Neves, R.P.P., Martins, S.A., Moorthy, N.S.H.N., Fernandes, P.A. and Ramos, M.J. (2013). Protein-ligand docking in the new millennium– a retrospective of 10 years in the field. *Curr Med Chem*, 20(18), 2296-2314.

Suresh, C. H., Vargheese, A. M., Vijayalakshmi, K. P., Mohan, N., & Koga, N. (2008). Role of structural water molecule in HIV protease-inhibitor complexes: A QM/MM study. *Journal of Computational Chemistry*, *29*(11), 1840-1849.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., & Jensen, L. J. (2014). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, *39*(suppl_1), D561-D568.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., & Kuhn, M. (2016). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, *43*(D1), D447-D452.

Toma, S., & Lafleur, L. (1974). Survey on the Incidence of Yersinia enterocolitica infection in canada. *Applied Microbiology*, *28*(3), 469-473.

Trends, E. F. S. A. (2009). Sources of zoonoses and zoonotic agents in the European Union in 2007.

Turner, p., Mclennan, A. & White, M. (2007). BIOS Instant Notes in Molecular Biology. Abingdon: Taylor and Francis.

Vaught, A. (1996). Graphing with Gnuplot and Xmgr: two graphing packages available under linux. *Linux Journal*, *1996*(28es), 7.

Wauters, G., Janssens, M., Steigerwalt, A. G., & Brenner, D. J. (1988). Yersinia mollaretii sp. nov. and Yersinia bercovieri sp. nov., formerly called Yersinia enterocolitica biogroups 3A and 3B. *International Journal of Systematic and Evolutionary Microbiology*, *38*(4), 424-429.

Wiederstein, M., & Sippl, M. J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research*, *35*(suppl_2), 407-410.

Wilde, R. E., & Singh, S. (1998). Statistical mechanics: Fundamentals and modern applications. *Interscience*.

Wu, S., Skolnick, J. & Zhang, Y. (2007). Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biology*, *5*(1), 17.

Yu, C. S., Chen, Y. C., Lu, C. H., & Hwang, J. K. (2006). Prediction of protein subcellular localization. *Proteins: Structure, Function, and Bioinformatics*, *64*(3), 643-

Yu, C. S., Cheng, C. W., Su, W. C., Chang, K. C., Huang, S. W., Hwang, J. K., & Lu, C. H. (2014). CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation. *PLoS One*, *9*(6), e99368.

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., & Brinkman, F. S. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, *26*(13), 1608-1615.

Zhang, R., & Lin, Y. (2008). DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Research*, *37*(suppl_1), D455-D458.