CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

Genome Based Drug Target Identification in Human Pathogen *Streptococcus gallolyticus*

by

Nosheen Afzal Qureshi

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

in the

Faculty of Health and Life Sciences
Department of Bioinformatics and Biosciences

2020

I dedicate this thesis to my parents and my teachers.

CERTIFICATE OF APPROVAL

Genome Based Drug Target Identification in Human Pathogen *Streptococcus gallolyticus*

by

Nosheen Afzal Qureshi

(MBS181013)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Uzma Abdullah	PMASAAU, RWP
(b)	Internal Examiner	Dr. Samra Bashir	CUST, Islamabad
(c)	Supervisor	Dr. Syeda Marriam Bakhtiar	CUST, Islamabad

Dr. Syeda Marriam Bakhtiar
Thesis Supervisor
August, 2020

Dr. Sahar Fazal
Head
Dept. of Biosciences & Bioinformatics
August, 2020

Dr. Muhammad Abdul Qadir
Dean
Faculty of Health & Life Sciences
August, 2020

Author's Declaration

I, **Nosheen Afzal Qureshi** hereby state that my MS thesis titled "**Genome Based Drug Target Identification in Human Pathogen *Streptococcus gallolyticus***" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

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

(Nosheen Afzal Qureshi)

Registration No: MBS181013

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Genome Based Drug Target Identification in Human Pathogen *Streptococcus gallolyticus***" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been dully acknowledged and that complete thesis has been written by me.

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

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

(Nosheen Afzal Qureshi)

Registration No: MBS181013

Acknowledgements

IIn the name of **Allah**, the Most Gracious and the Mot Merciful Alhamdulillah, all praises to Allah for giving me strength and for His blessings in completing my MS thesis. First, I would like to express my sincere gratitude to Capital University of Science and Technology (CUST) Islamabad for providing me and opportunity to do MS Biosciences and achieving my goal to pursue higher studies. I would like to start with a special appreciation that goes to my Supervisor, **Dr. Syeda Marriam Bakhtiar**, for her constant support, encouragement and guidance throughout this thesis. her door was always open whenever I needed help, she always guided me as a mentor. Then I would like to pay my special thanks to my CO-Supervisor, Syed Babar Jamal Bacha for his constant support and motivation throughout the process of this thesis. I would like to thanks to my teachers Dr.Shaukat Iqbal, Dr.Erum Dilshad, Dr. Arshia Amin and Dr. Sahar Fazal. Special thanks to my friends and colleagues for supporting me throughout this time. Finally, I express my gratitude to my parents and siblings for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis.

(Nosheen Afzal Qureshi)
Registration No: MBS181013

Abstract

Streptococcus gallolysticus (Sg) previously known as *Streptococcus bovis*, is Gram positive, non-motile bacteria. This bacterium is known to cause infective endocarditis which is an inflammation of inner lining of heart. As treatment of this disease is quite expensive and some of antibiotics against this disease have already shown resistance, therefore, it is vital to find the novel therapeutic targets and potent drugs to prevent the onset of this disease. In this study, we have used in-silico approach to link genomic data of *Sg* species with its proteome for identification of putative therapeutic targets. We have identified 1,138 core proteins by using pan genomic approach. Further, using subtractive proteomic analysis a set of 18 proteins was selected as these targets were essential for bacteria and non-homologous to host (human). The drug prioritization allows us to identify the drug and vaccine targets. The selected proteins were subjected to molecular docking against drug like compounds retrieved from zinc library. Furthermore, the best dock compounds with lower binding energy were identified. In this work we have identified a novel drugs/vaccine targets against *Sg*, of which, some have already been reported and validated in another species. Owning to the experimental validation we believe our methodology and result are positive contribution for drug/vaccine target identification against *Sg* caused infective endocarditis.

Keywords: *Streptococcus gallollyticus*, Infective endocarditis, Pangeome, Subtractive proteomic analysis.

Contents

List of Figures

List of Tables

Abbreviations

Blast	Basic Local Alignment search tool
DEG	Database of Essential Genes
Edgar	Efficient Database framework for comparative Genome Analyses using BLAST score Ratios
Kegg	Kyoto Encyclopedia of Genes and Genomes
MOE	Molecular Operating Environment
MW	Molecular Weight
Sg	*Streptococcus gallolyticus*
Uniprot	Universal Protein Resource
Vfdb	Virulence factors Database

Chapter 1

Introduction

Streptococcus gallolyticus (Sg) is Gram positive, non-motile bacteria. *Sg* previously known as *S. bovis* which is phenotypically diverse bacteria belonging to Lancified Group D Streptococci [1]. This bacterium is an opportunistic pathogen causing many diseases such as infective endocarditis, colon cancer, meningitis and septicemia. For many years the classification of *Sg* has been revised several times [2]. Previously *S. bovis* was classified as three biotypes, biotype-I belonging to *Sg subsp. gallolyticus*, biotype-II/1 belonging to *Streptococcus infantarius subsp. infantarius* and *Streptococcus infantarius subsp. coli* and biotype-II/2 belonging to *Sg subsp. pasteurianus* [3] and currently on the basis of multilocus sequence typing, the classification is divided into 7 subspecies which are; *Sg subsp. gallolyticus, Sg subsp. macedonicus, Sg subsp. pasteurianus, Streptococcus infantarius subsp. infantarius, Streptococcus lutetiensis, Streptococcus alactolyticus and Streptococcus equins* [1]. This bacterium grows in chain or pairs and is non-γ-hemolytic or slightly γ-hemolytic but sometimes shows alpha-hemolytic activity on ovine blood agar plates [4], [5]. It is commonly present in microflora and appears approximately 2.5-15% in gastrointestinal tract of healthy individual [6].

Rusniok et al. (2010) [3] completed the whole genome sequence of *Sg* which was isolated from patient suffering from infective endocarditis. They identified some virulence factors particularly involve in causing disease and some metabolic

characters which provides the organism the ability to utilize various types of carbohydrates in gut microbiota which increases the survival of the pathogen in an organism. The gene encoding virulence factors identified by Rusniok el al. (2010) was likely involve in polysaccharide production, glucan mucopolysaccharide which is a putative component of biofilm produce by this species, 3 types of pili and collagen binding protein [3]. This production of polysaccharides provides protection to *Sg* innate immune response produce by host. The glucan mucopolysaccharide, collagen binding proteins and 3 types of pilli (pil1,pil2 and pil3) helps in adherence to endothelium cells which is the initial step in the development of infective endocarditis [5]. There was also some evidence which showed that some of these genes may be acquires by horizontal gene transfer from other gut species. The proteins which are produce by these encoding genes not only increase the survival of this bacteria in harsh environment but also make it more pathogenic in causing disease [3].

This bacterium is opportunistic human pathogen which causes endocarditis disease, a serious infection of inner lining of the heart [7]. For the last few years, a significant increase in amount of cases of infective endocarditis were observed [8]. Hoen el al. documented that a significant proportion of streptococcal infective endocarditis cases is responsible for *Sg*: 58% in France, 9.4% in other European countries and 16.7% in the USA. [9]. This disease mostly occurs in elderly patients [10]. The recent survey in developed countries shows that among 100,000 population 2.6 to 7 cases of endocarditis has been reported per year. The median age of these patients were 58 or greater than 58 [11]. The risk of developing *Sg* endocarditis could be uncooked meat or fresh dairy products or whose immune system is weak. The patient who have associated hepatic diseases have high rate of mobility and mortality disease. Severe infection has been reported in those elderly patients who has co-morbities such as diabetes mellitus, hepatic disease, rheumatic disorders [12]. The risk factor for developing infective endocarditis are Age over 60, male gender, use of injection medicine, history of prior infectious endocarditis, weak dentition or dental treatment, involvement of a prosthetic valve

or intracardiac device, history of valvular disease such as rheumatic heart disease, aortic valve disease and congenital heart disease such as pulmonary stenoid [11].

Boleij et al. (2011) [4] identified different type of virulence characteristics of *Sg* which might play an important role in pathogenesis of infective endocarditis in an organism. The researcher used different types of human epithelial tumor cells to identify the pathogenesis of *Sg*. This research was in vitro study in which they used series of assay which were based on cell responses produce by both bacteria and human. These studies focused on *Sg* adhesion, invasion, translocation, development of biofilms and ability to induce an immune-response to identify virulence characteristics that could explain the relation between *Sg* infective endocarditis and colon carcinoma. The findings of this work indicate that this bacterium is remarkably capable of translocating through malignant intestinal epithelium in a paracellular fashion without inducing a major immune-response and then adhering to collagen-rich surfaces and forming biofilms. Such results are interesting in that they discuss the main putative first steps in the propagation of bacteria from the gut to the bloodstream and eventually to other more colonization-friendly sites, such as collagen-rich cardiac valve surfaces[4].

The first step in pathogenesis of infective endocarditis caused by *Sg* is endocardial injury. The injury occurs due to turbulent flow of blood on valve surface typically on atrial surface of atrioventricular valves or ventricular surface of semilunar valves. The endocardial injury triggers the thrombus formation which due to the removal of fibrin and platelets. After thrombus formation the bacteria enters into the bloodstream through thrombus. As *Sg* has virulence properties proposed by Boleij and colleague, it enters into the bloodstream in a paracellular manner without inducing major immune response and adheres to the damage collagen rich surface of cardiac valve (endocardium). Once it attached to the endocardium, this bacterium proliferates and forms biofilm which causes the inflammation in the lining of heart and causes endocarditis [4], [13]. For a generalist, endocarditis is seldom an easy diagnosis. It may present with a wide variety of clinical signs,

some subtle; diagnosis may be difficult or indications may be deceptive, and consideration should be given to a wide differential diagnosis. The clinical symptoms of infective endocarditis are described in figure 1.1 [14].

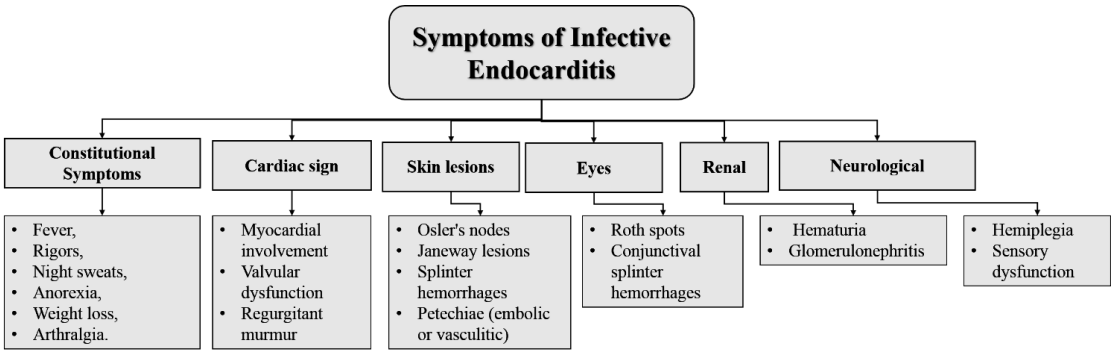

FIGURE 1.1: Symptoms of Infective Endocarditis [14]

Sg is resistance to penicillin so the preferred medication treatment for infective endocarditis is penicillin G with gentamycin and estreptomicin. Certain options could be Gentamicin-related Ceftriaxone in one daily dose. If patient is allergic to penicillin then vancomycin is preferred[15]. For patients with persistent fever resistant to medical therapy, surgical intervention may be needed. Surgery is also recommended for those who have valve obstruction, mitral regurgitation, paravalvular abscess, production of the Valsalva sinus aneurysm, multiple embolic episodes, gradual cardiac insufficiency with serious valve damage and oscillating vegetation of >1cm. Surgery can require aortic root replacement for aortic root abscesses, as well as valve replacement. A full course of antibiotic eradication therapy should be given after the relevant surgical procedure [14]. The cost of the surgery for endocarditis is very expensive [16] and one of the strains of *Sg* is also found to be resistance to tetracycline [6]. So, there is need to identify novel therapeutic targets and potent drugs to prevent the onset of disease.

For this, the current study was designed to integrate in-silico approaches to link genomic data of *Sg* species with its proteome for identification of putative therapeutic targets. It can be used to classify potent inhibitors that may contribute to the discovery of compounds that inhibit pathogenic development. The proteomes from the 7 genomes of *Sg* were compared using pangenome approach [17]. The entire gene set of all strains of species is called pan genome. Its incudes core genome

which the genes present in all strains of species, accessory genome which are the genes present in two or more strains but not in all of them and the singletons restricted to only one strain of species. It provides the genomic diversity present between the strains of a distinct species [6]. Then predicted core genome is further filtered out on the basis of essentiality for the bacteria. Then using subtractive genomic approach all essential proteins were check for the non-homologous to the host (human) and then all non-host homologous protein were subjected to virtual screening using compound library of 11,993 retrieved from zinc database. The putative targets that were identified might be used to design peptide vaccines and suggest novel lead, natural and drug-like compounds that could bind to the proposed target proteins [17].

1.1 Problem Statement

Streptococcus gallolyticus causes infective endocarditis which is disease of an inner lining of heart. For many years, this disease is being treated with antibodies but recent research on this disease have shown antibiotic resistance against one of the strains of *Streptococcus gallolyticus*.

1.2 Proposed Solution

There is a need of alternative novel targets and potent therapeutics to prevent the onset of the disease.

1.3 Aim of Study

The aim of the study is to identify novel and potent therapeutics targets to prevent the onset of the infective endocarditis disease by using pan-genomic and subtractive genomic approach approach.

1.4 Objectives

The objectives of this study include:

1. To identify core genomes of all strains of *Streptococcus gallolyticus*.
2. To perform subtractive genomics analysis.
3. To prioritize our protein targets, identification of potent lead compound using protein-ligand interaction.

1.5 Scope

Bioinformatics, in general, contributes through prediction of therapeutic targets which ultimately reduce men efforts and cost of experimentation. So, in this study, we will contribute towards drug development against endocarditis disease by predicting novel therapeutic targets and potent lead compound for inhibition of identified targets. The promising ligand molecule can be tested in experimental laboratory that can ultimately result in commercial product in future.

Chapter 2

Literature Review

2.1 Background

Streptococcus gallolyticus is Gram positive bacteria which an opportunistic pathogen in causing infective endocarditis (IE) [3]. Rusniok *et al* . (2010) also found some virulence genes which play important role in causing this disease. Infective endocarditis is an infection of inner lining of heart. This disease is more common in man than women. This disease is very prevalent in western countries as well in the Asian countries but the difference is that in western countries this disease is more prevalent in elder people but in Asian countries such as Pakistan, India and China this disease is more common in younger people age ranges from 34-40 years. This difference is mainly because of the rheumatic heart disease is more prevalent in Asian countries. In Pakistan, the person who is suffering from rheumatic heart disease is likely to have infective endocarditis. The main issue in treating this disease is the antibiotic resistance, so, in this study, we have used pangenomic and subtractive genomic approach to find the new and common drug targets for all the strains of Sg to overcome the antibiotic resistance.

2.2 *Streptococcus gallolyticus*

Streptococcus gallolyticus (Sg) previously known as *S. bovis*, a phenotypically diverse bacteria belonging to Lancified Group D *Streptococci* [1]. This bacterium is non-motile, grows in chain or pairs and is non-γ-hemolytic or minimally γ-hemolytic but sometimes shows alpha-hemolytic activity on ovine blood agar plates [4], [5]. It is a common member of microflora and appears approximately 2.5-15% in gastrointestinal tract of healthy individual [6]. This bacterium is an opportunistic pathogen causing many diseases such as infective endocarditis, colon cancer, meningitis and septicemia [2]. Rusniok *et al.* (2010) [18] completed the whole genome sequence of *Sg*. This strain was isolated from patient who was suffering from infective endocarditis and colon cancer. They found the genes encoding proteins and enzymes that could play an important role in survival advantage in gut and may involve in causing the disease. The existence of genes encoding enzymes that may be involved in the digestion of plant cell wall polysaccharides and tannins (a toxic by-product of which is gallate, which this bacterium is uniquely capable of using as a source of carbon) and biosynthetic pathways for pantothenate, nicotinamide adenine dinucleotide, and glutamate were unique to streptococcal strains. Such metabolic traits are likely to provide the capacity of this organism to use different carbohydrates in the gut and give it a clear advantage over other species in survival. In addition, this study also identified numerous genes encoding potential virulence factors, including genes likely to produce a polysaccharide capsule, glucan mucopolysaccharides (including hemicellulose, a putative biofilm component produced by this species), 3 types of pili, and collagen-binding proteins. Furthermore, there is evidence that some of these genes may have been acquired from other gut organisms by lateral gene transfer. The proteins encoded by these genes may not only allow the organism to create a niche in the harsh intestinal environment, but may also become invasive and cause endovascular infection. Data showing that *Sg* isolates from infective endocarditis patients express collagen-binding adhesives and pilus proteins [18] and adheres to extracellular matrix proteins found in aortic valves, including collagen types I and IV [19], as well as endothelial cells, are

consistent with the findings from genome analysis [5], [6]. The biofilm formation with the host cell has been shown in figure 2.1.

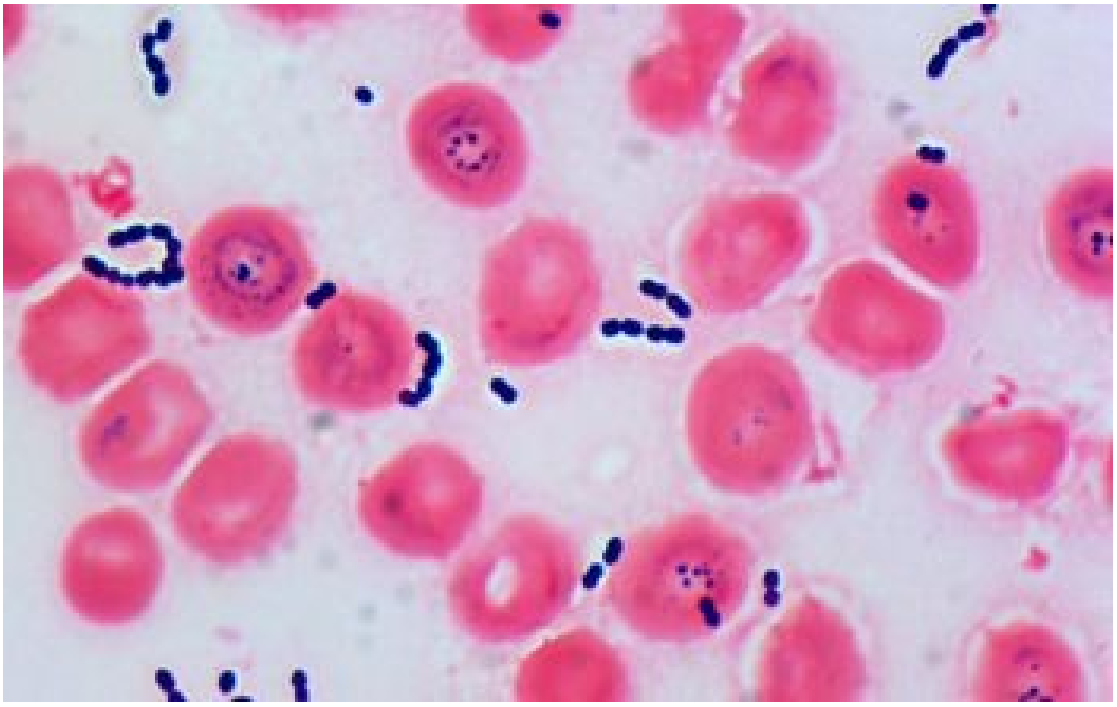

FIGURE 2.1: Biofilm Formation of *Streptococcus gallolyticus*. [20]

2.3 Infective Endocarditis

Infective endocarditis [IE] is a disease which is an inflammation of inner inning of heart. Yearly rate of this is quite low from 3 to 7 people are infected per year but this disease is characterized as most lethal and life-threatening disease due to its high morbidity and mortality rate. Globally, in 2010, IE was associated with 1.58 million disability-adjusted life-years or years of healthy life lost as a result of death and non-fatal illness [21]. This disease is mostly occurring in elderly patient due to their weak immune system[12].

Infective endocarditis is known to be caused by *Sg* [4]. Boleij *et al.* (2011) [4] identified several virulence characteristics of *Sg* which may play important role in pathogenesis of infective endocarditis in an organism. These researchers used differentiated human epithelial colorectal adenocarcinoma cell to replicate the path

of the *Sg* infection in vitro in a series of assay based on both bacterial and host cell responses. These studies focused on *Sg* adhesion, invasion, translocation, development of biofilms and ability to evoke an immune response to identify virulence characteristics that could explain the relation between *Sg* infective endocarditis and colon carcinoma.

The findings of this work indicate that this bacterium is exceptionally capable of translocating through malignant intestinal epithelium in a paracellular fashion without inducing a major immune response and then adhering to collagen-rich surfaces and forming biofilms. These results are interesting in that they discuss the main putative first steps in the propagation of bacteria from the gut to the bloodstream and eventually to other more colonization-friendly sites, such as collagen-rich cardiac valve surfaces [4].

Hinse *et al.* (2011) [6] sequence the one of strains of *Sg* and identified some surface proteins, virulence factors and protective elements which may play role in pathogenicity and adhesion to endocardium. They also identified the p*Sg*G1 plasmid which contained 21 ORFs including tetracycline resistance genes and which may play a functional role in horizontal gene transfer in an organism [6]. Isenring *et al.* (2018) [21] identified novel pathway for causing endocarditis. They proposed steps that how it causes the infective endocarditis.

First it survives in human blood after entering into the blood stream. Then it activates the cellular components of coagulation/clotting cascade and induction of procoagulant state. Pil1 helps *Sg* to adhere on the collagen rich surface of the heart and with release of bradykinin and its binding to its receptor B2R triggers the infective endocarditis. Bradykinin helps recruitment of monocytes and neutrophils boosting up the activation process of neutrophils in human body.

Interfering with contact activation would be an attractive target for treatment but this study is still understudy whether these events take place in vivo or not. Their data helps us to understand the pathogenesis of *Sg* causing infective endocarditis and also highlighted the virulent genes taking part in causing the disease Pathogenesis of infective endocarditis by *Sg* is shown in figure 2.2.[22].

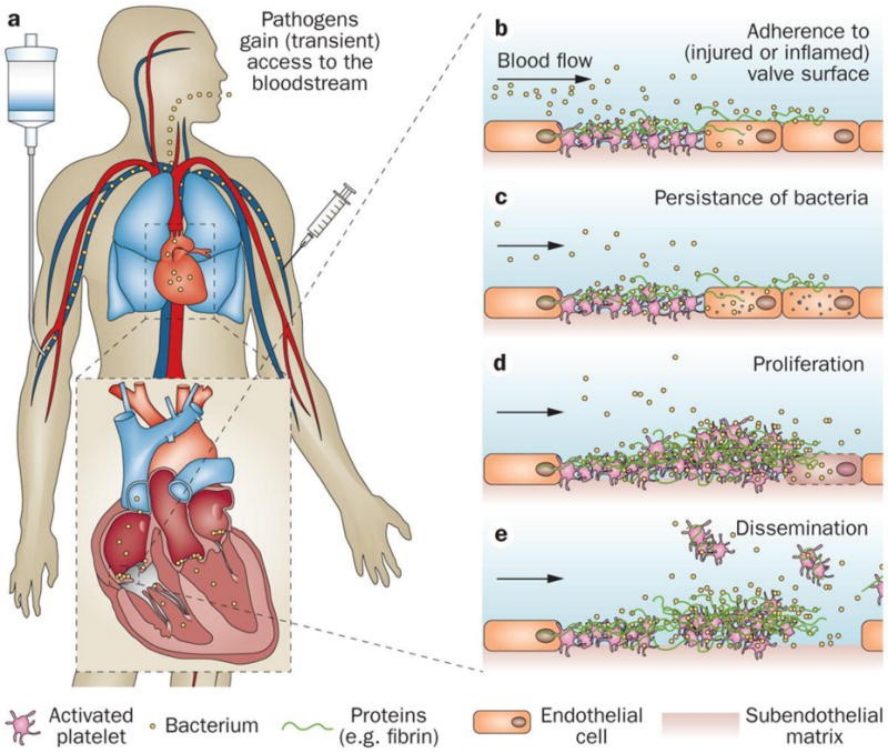

FIGURE 2.2: Pathogenesis of infective endocarditis by *Sg* (host-pathogen interaction) [23] a) Bacteria enters into blood stream via injection or intravenous catheter b) adherence of *Sg* to the collagen rich surface c) gain access to valve endothelium d) proliferation of *Sg* e) disseminate in the form of emboli which could lead to myocotic, ischemic stroke and abscesses

2.4 Prevalence of Infective Endocarditis in Pakistan

Infective endocarditis is an important cause of morbidity and mortality in Pakistan. In our country, infective endocarditis occurs at lower age. Many patients who are suffering from this disease have age less than 40. In 2004 Tariq *et al.* reported that median age of infective endocarditis was 24 years [24], [25] and in 2015 same group reported a shift in the median age which was 34 years. This gradual shift could be explained due to shift from communicable to non-communicable disease. The non-communicable disease includes cardiovascular disease, diabetes, cancers and chronic airways diseases. Proper medical care is required in managing this disease [26]. These results are quite different from developed countries, as in developed countries (western countries) the age of infective endocarditis is

greater than 50 [11]. In our country the infective endocarditis occurring at lower age is due to high frequency of rheumatic heart disease, multifactorial diseases and unrepaired congenital heart disease. The studies have shown that men are more effected than female from infected endocarditis [27]. The ratio of infective endocarditis in men and women is 2:1 respectively[25]. The high prevalence of this disease in men is because men has more access to medical care and thus exposed to nosocomial infections and intravenous drug usage is also very common in men [27]. Another reason for this was explained by Durente el al. that male predominance of infective endocarditis decreases with the age and female hormones play a protective role against infective endocarditis [28] that's why it is more common in men than in females. From the blood and tissue culture of infected patients the most frequently organism isolated is streptococcus group (36.7%) in which about 14.2% are susceptible to penicillin. These findings are quite similar from the neighboring countries like China and India [26]. Among the reported cases about 5% of patients are suffering from infective endocarditis caused by *Sg* [29]. These results show that in Pakistan the frequent group causing infective endocarditis is streptococcus group.

2.5 Pan-Genome

The pan-genome is the entire gene set that are present in given dataset. It includes core genome that are those genes which are shared by all genomes, accessory genome that those genes which are absent in some of the strains and then strain specific genes (singletons) or unique genes which are only present in single genome. The figure 2.3 shows the selection criteria of pangenome. In this the gene set of all three strains are collectively called pan-genome, the red color shows the core genome which are those genes shared by all genomes, the yellow color shows accessory genome in which are those genes that are absent in some of the strains and orange, blue and green color shows the strain specific or unique gene which are only present in single genome.[30]. It provides the genomic diversity present between the strains of a distinct species [6]. In the case of *Streptococcus agalactiae*,

Tettelin *et al.* (2005) [23] was the first to describe the pan-genome concept and they also showed the core genome of eight strains consist of 80% of genes of any single strain and when the data analysis was done it shows enormous gene pool [31]. This could give us the hint that larger number of sequenced strains could provide more information about genomic diversity present in distinct species [6]. Here is some tool mentioned in table 2.1 which are available for pan genome analysis.

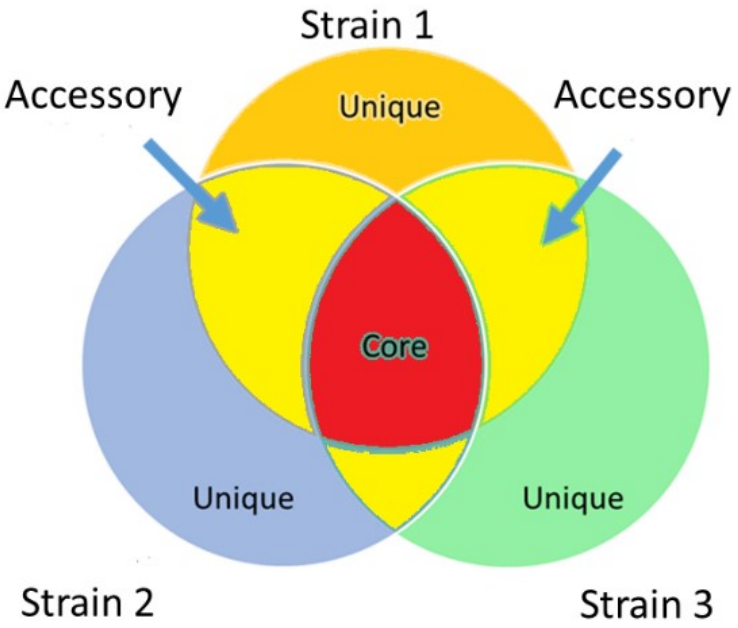

FIGURE 2.3: Core Genome Selection Criteria from Pan-genome.[32]

TABLE 2.1: List of Some Available Tool for Pan Genome Analysis Along with Their Functions.

Tool	URLs	Function	Ref.
EDGAR	edgar. computational. bio.uni-giessen. de	This tool does the homology analyses in which data set is automatically adjusted according to the given data.	[33]

PGAT (Prokaryotic Genome Analysis Tool)	nwrce.org/pgat	This tool compares the multiple strains of same species and predicts the genetic data of the species.	[30]
PGAP-Pan-genome Analysis Pipeline	http:// pgap.sf.net	This tool does the pan-genome analyses, functional analyses of collection of genes and genes evolution and genetic variation.	[30]
PanGP	http://PanGP.big.ac.cn	This tool is time efficient tool as it does the pangenome analysis for large scale genomes at very less time.	[30]
ITEP-Integrated Toolkit for the Exploration of Microbial Pan-genomes	https://price.systems biology.net/itep	This tool predicts the genetic variation of the protein, find its orthologous protein, pangenome analysis and find the metabolic networks for related species.	[30]

GET_ HOMOLOGUES	http://www.eead. csic.es/compbio/ soft/gethoms.php	This tool does the comparative genomic analysis and pan genome analysis of the bacterial strains.	[30]
Panseq– Pan-genome Sequence Analysis Program	http://76.70 .11.198/ panseq	This tool predicts the core and accessory genome on the basis of sequence identity and segmentation length. It does not predict the core genome on the basis of proteins.	[30]
OrthoMCL	http://www.cbil. upenn.edu	This tool uses Markov cluster algorithm to construct the orthologous group and its paralog of multiple eukaryotic species.	[34]

In this study, EDGAR tool was used. This web tool is fast and user-friendly. It provides user-friendly survey of evolutionary relations between the bacterial species and the process for obtaining new biological insight into different gene content is quite easy and understandable. It allows easy browsing by providing all features in web based and also private account which is the platform-independent user interface [33].

TABLE 2.2: List of Publication Using Pan Genome Based in Identification of Drug Targets and Vaccine Development

Target Bacteria	Analysis	Outcomes	Ref.
Clostridium botulinum	To understand the symptoms of pathogen pan genome analysis was done. For prediction of specific targets such as drugs and vaccine targets core genome analysis were done.	They found high genomic similarity from the genomes. Found those set of genes which are virulent can cause threat to the human population.	[35]
Helicobacter pylori	They have used pangenome and reverse vaccinology approach to identify the potential core immunogenic genes.	They found 1,193 conserved genes out of which 28 were non host homolog proteins that could act as therapeutic targets against H.pylori.	[36]
Corynebacterium diphtheria	Pan-genomic approach to identify the to identify the putative therapeutics targets.	Identified 23 conserved targets out of which 8 protein were essential and non-homolog and these protein shows the therapeutic properties for drugs and vaccine development.	[17]

Corynebacterium pseudotuberculosis	MHOLline workflow (Pan-modelome and core modelome)· Subtractive genomic approch	Identified 10 conserved proteins out of which only 4 were essential and non-homolog which shows favorable interaction with top ranking compounds.	[37]
Treponema pallidum	MHOLline workflow (Pan-modelome and core modelome) Reverse vaccinology and Subtractive genomic	6 non homolog and essential proteins were identified which were also the conserved protein which could be for vaccine and drug development against Treponema pallidum.	[38]

2.6 Subtractive Genomic Analysis

Subtractive genomics is the mechanism by which sequences between the host and the pathogen proteome are subtracted, which helps to provide data for a collection of proteins that are important for pathogen but not present in the host [39]. Essential genes are the gene that helps an organism to survive. Deletion of these genes could cause cell death which indicates that these genes are involve in essential biological function [40]. The Table 2.3 shows the list of available tools for identification of essential genes.

TABLE 2.3: List of Some Available Tools for Identifying Essential Genes

Name of tool	Description	URLs	Ref.
DEG	This database consists of essential and non-essential data of archaea, prokaryotes & eukaryotes.	http://origin.tubic.org/deg/public/index.php	[41]
Gene Essentiality database (OGEE)	This is online database which provide data about essential and non-essential genes from large based experiments. This database is mainly focused on Cancer related essential genes	http://ogee.medgenius.info/browse/	[41]
Essential Genes on Genome Scale (EGGS)	This database consists of microbial gene essentiality data.	http://www.nmpdr.org/FIG/eggs.cgi	[41]
CEG	To identify essential genes this database uses CEG_MATCH tool. This database is link with DEG database and uses its data and shows the gene clusters.	http://cefg.uestc.edu.cn/ceg	[41], [42]

EGP-Essential Gene Prediction	It is an online tool for identification of essential genes of bacterial genome by using SVM based method. The accuracy of this tool is quite low.	http://cefg. uestc.edu. cn:9999/egp	[41]

In this study, Database of Essential Gene (DEG) (http://origin.tubic.org/deg/public/index.php)was used . This database provides large set of essential genes from which the user could easily blast the query sequence genes against the genes present in DEG. For each query gene it gives unique DEG identification number, the gene name, its function and sequence [43]. The next step is to identify non-host homologous proteins. Non-host homolog proteins are those protein which not present in the host but present in the pathogen. Non-homologous proteins were identified by Blastp tool (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) which helps us to filter out those essential proteins which are not present in host (human).

2.7 Drug Target Prioritization

Drug prioritization includes many factors such as molecular weight, functionality, Subcellular localization, pathway analysis and virulent genes. Each step is important for determining the putative drug and vaccines targets.

2.7.1 Molecular Weight

Molecular weight is the weight of the protein which is calculated in Kilo Dalton (kDa). In terms of drug target, the molecular weight of the target protein

should be less than 100 kDa because it could easily pass through the cell membrane and reached its target site[44]. The molecular weight was calculated from ProtParam tool. ProtParam is an online tool that is available on Expasy server. This tool computes physio chemical properties of the molecule. The properties include molecular weight, iso-electric point, amino acid composition and atomic composition etc[45].

2.7.2 Functionality

The functionality of the drug/vaccine targets is determined by their role in molecular and biological processes. Molecular function is defined as all activities that taking place at molecular level. It usually corresponds to those activities that are performed by individual gene products such as protein or RNA but some activities are performed by molecular complexes. These molecular complexes consist of multiple gene products. The biological process includes those activities which is performed by multiple molecular activities such as DNA repair[46]. In this study the molecular and biological processes was performed by Uniprot. Uniprot (Universal Protein Resource) is an online freely available database of proteins which contains collection of protein sequences along with their functional information. Uniport association is collaboration of EBI (European Bioinformatics institute, PIR (the Protein Information Resource) and SIB (Swiss Institute of bioinformatics. This database consists of large amount of data about their biological and molecular function of proteins. This database provides user-friendly platform to achieve certain task and it is updated every three weeks so this database could be quite valid for this study[47].

2.7.3 Subcellular Localization

It is very important to find subcellular location of proteins as it provides information about the environment in which they perform their activities. Subcellular

localization is capable to influence any protein function by controlling the accessibility and availability of all types of molecular interaction partners. The knowledge of protein localization plays a significant role in characterizing cellular function of newly discovered protein or hypothetical proteins [48]. There number of online web-based tools through which could determine the cellular localization of the protein. Some of them are mentioned in a table 2.4.

TABLE 2.4: List of Some Online Available Tool for Subcellular Localization

Name	Description	URLs	Ref.
CELLO	This tool is based on peptide composition to predict the subcellular localization of the protein.	http://cello.life.nctu.edu.tw.	[49]
TargetP	It is neural network based model which is use for the prediction of subcellular localization in Eukaryotes	http://www.cbs.dtu.dk/services/TargetP/	[51]
SubLoc	SubLoc is based on Support vector machine method which use to predict the subcellular localization in prokaryotes.	http://www.bioinfo.tsinghua.edu.cn/SubLoc/.	[52]

PSORT B	It is a web-based application which is used to predict the localization, transmembrane alpha-helices and motifs for Gram negative bacteria.	http://www.psort.org	[50]

In this study I have used CELLO web-based tool because it has better accuracy rate among all the above mention subcellular localization tool. This is simple straight forward tool based on Support Vector Machine (SVM). When this tool was compared with PsortB tool, Cello give much better result that PsortB. This tool is quite simple and user friendly. This tool could be useful for high throughput and large-scale analysis of genomic and proteomic data[49].

2.7.4 Pathway Analysis

Pathway analysis plays an important role in term of drug discovery. National Human Genome Research institute describe biological pathway as a series of action among molecules in a cell that leads to certain product or change in the cell. In terms of function the biological pathway are classified in three categories metabolic pathway which is involve in many chemical reaction in biosynthesis or decomposition of many metabolites, Gene regulation which is responsible for on and off genetic information flow that predicts the protein expression and Signal transduction pathway which are responsible for carrying signal from externa environment to interior cellular compartments [53]. In this study KEGG (Kyoto Encyclopedia of Genes and Genome) was used for Pathway analysis. This is widely used database for pathway analysis. It consists of manually drawn reference pathway

along with organism-specific pathway that is computationally generated by matching KO assignments in the genome with references pathway maps. It is quite user friendly and much reliable database in term for pathway analysis[54].

2.7.5 Identification of Virulence Genes

To identify the therapeutic targets identification of virulent genes is very important because these gene play a key role in causing the disease, as it the ability of pathogen to cause disease. They could be involved in adhesion, invasion, colonization's, in producing toxins to damage tissues of host, capsular polysaccharide or siderophores which takes up the iron. So, identification of these genes is very essential and could be used as potential drug target for treating infectious disease [55]. List of available tools for identification of virulence genes is shown in table 2.5.

TABLE 2.5: List of Some Online Available Tool for Identifying Virulent Genes

Name	Description	URL's	Ref.
VFDB	It is an integrated and comprehensive database which contains data about virulence factor of pathogenic bacteria.	http://www.mgc.ac.cn/VFs/	[56]
VirulentPred	This tool identifies virulent factors using SVM based two layer method.	http://203.92.44.117/virulent/	[56]

MvirDB	It uses microbial annotation database to maintain its data. It includes data about virulence factors, its toxicity and antibiotic resistance. This server has not updated since 2007.	http://mvirdb .llnl.gov/	[56]
GI-POP	It is a web-based tool which predicts data about genomic island, genome assemblies and	http://gipop. life.nthu.edu.tw/	[56]
Islander	This tool identifies genomic island of bacterial genome.	http://bioinf. iiit.ac.in/IGIPT/	[56]
PAIDB	It is comprehensive database for identifying the pathogenic island. This database has not been updated since 2008.		

For the identification of virulent genes Virulence Factor of Database (VFDB) was used. It is the web-based virulence factor database which is focuses on human

bacterial pathogens. It provides user the rapid access to current knowledge of virulence factor from various bacterial pathogens. It consists of large amount of data approximately 3224 sequences. It is quite understandable and user friendly [57].

2.7.6 Catalytic Pocket Detection

The detection of catalytic pocket is quite essential as they represent the targets for low molecular drug [58]. For the catalytic pocket detection DoGSiteScorer was used. This tool is automated pocket detection and analysis tool which is used for calculation of druggability of protein cavities. This tool will return the pocket residue and druggability score which ranges from 0-1. The score closer to 1 indicates highly druggable protein cavity. The predicted cavities are likely to bind ligands with high affinity [17].

2.7.7 Retrieval of Ligands

The ligands against *Sg* were retrieved by reviewing literature. There are number of tools to download the structure of ligands such as Zinc database, Pubchem, Chembl, NCI, BindingDB and PDB-Bind [59]. In this study, Zinc Database was used. [60].

2.7.8 Molecular Docking

Molecular docking is the process which is used to understand the interaction between the drug and molecule for new drug discovery. In this method the protein binding site/active site interact with the ligands in non-covalent manner in forming a stable complex with more accurately and effectively [61]. There are many tools which are available for molecular docking to analyze the protein-ligand interaction some of them are listed in table 2.6. In this study MOE tool was used.

TABLE 2.6: List of Some Available Tools for Molecular Docking

Name	Description	URL's	Ref
AutoDock	It is a ligand-protein modelling tool for drug designing. This is desktop-based tool.	http://autodock.scripps.edu/	[62]
GOLD	It is combined mechanistic tool which includes molecular docking and quantum mechanics to model the protein-ligand interaction. It is a commercialized-based tool.	https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold/	[62]
MOE-Molecular operating Environment	This tool is a molecular modeler tool and user-friendly tool. This tool works on Window and linux.	https://www.chemcomp.com/Products.htm	[63]
GLIDE	It is commercialized based tool which is used for protein-ligand interaction.	https://www.schrodinger.com/glide	[62]

Chapter 3

Material and Methods

All complete strains of *Sg* were selected to identify the core genomes using pan genomic approach. Then selected core genomes were further narrow down by using subtractive genome which includes all those genes which are non-host homologous and essential for the survival of bacteria. Then drug prioritization and protein-ligand interaction were done which leads us to novel targets and potent therapeutics to prevent the onset of the disease.

3.1 Genome Selection

The seven complete genomes of *Sg* were included in this study (Table 3.1). All the gene and protein sequences of this bacterium were retrieved from NCBI (https://www.ncbi.nlm.nih.gov/genome/).

Table 3.1: Strains of Streptococcus gallolyticus with information on genome statistics and region of isolations

Strain	Genome size	GC %	Total genes	Total protein	Region
DSM 16831	2.4929	37.70	2498	2341	Australia
NCTC13773	2.49358	37.70	2496	2333	Australia
ATCC 43143	2.36224	37.50	2357	2229	-

ATCC BAA-2069	2.37721	37.60	2377	2218	Germany
UCN34	2.35091	37.60	2345	2215	-
ICDDRB-NRC-S1	2.0525	37.70	2125	1759	Bangladesh
NCTC8133	1.86767	37.50	1845	1733	-

3.2 Identification of Core Genomes

The pan-genome is the total number of genes that are present in given dataset. It includes core genome, accessory genome and strain specific genome. Core genome includes all of the genes that are shared by all genomes, accessory genome consist of those gene which are absent in some of the strains and strain-specific genes consist of those genes which are only present in single genome[30].

The core genome was identified using EDGAR software [33]. In this, among all, one strain is selected as a reference strain which could act as a template strain and rest of all the strains are compared with the reference strains and then only those core genomes are selected which are common in all the strains. The algorithm that it used was BLASTp with the standard scoring matrix BLOSUM62 and cut off value of E= 1x 10-5 [64].

3.3 Identification of Essential Genes

Essential genes are the genes which are important for the species to survive or involve in growth. Subtractive genomics approach was used for the selection of conserved target which are essential to *Sg*. The list of conserved proteins of *Sg* which were retrieved from EDGAR software was subjected to the database of essential genes: DEG (http://origin.tubic.org/deg/public/index.php).

Database of essential genes consist of experimentally validated data from eukaryotes, archaea and prokaryotes. To identify the essential genes from DEG default parameters were selected, e-value=0.001, identity≥35%, scoring matrix was BLOSUM62 [65].

3.4 Identification of Non-Host Homologous Proteins

Subtractive genomics is the mechanism by which sequences between the host and the pathogen proteome are subtracted, which helps to provide data for a collection of proteins that are important for pathogen but not present in the host [39]. The identification of non- host homologous and homologous protein was carried out using NCBI BLASTp, default parameter was used. e-value=0.0001, bit score ≥100, scoring matrix BLOSUM62 and identity ≥25% [17].

3.5 Drug Target Prioritization

There are several factors that can help in determining the potential therapeutic targets such as molecular weight, molecular function, cellular localization, pathway analysis and virulence [66]. Molecular weight was determined by ProtParam tool (http://web.expasy.org/protparam/).

Determination of molecular weight (MW) is very important for the drug target. The targets whose MW is <100 kiloDalton (kDa) they are considered as best therapeutic target [67]. Molecular functions and biological process for each protein target was determined by Uniprot(https://www .uniprot.org/). Subcellular localization of pathogen was performed by CELLO(http ://cello.life.nct u.edu.tw/).

The cellular localization of bacteria determines the environment in which proteins operate. It effects the function of protein by controlling accessibility and availability of all types of molecular interaction partners. The knowledge of protein

localization often plays an important role in characterizing the cellular function of hypothetical and newly discovered proteins[48]. For pathway analysis the KEGG web tool (https://www.genome.jp/kegg/) was used which was used to determine the role of protein targets in different cellular pathways. To identify virulence of protein targets VFDB(http://www.mgc.ac.cn/VFs/) was used which determine the pathogenic virulence of the protein targets.

3.6 Catalytic Pocket Detection

For the catalytic pocket detection DoGSiteScorer was used. It is an automated pocket detection tool which is used for calculation of druggability of protein cavities. This tool returns the pocket residue and druggability score which ranges from 0-1. The score closer to 1 is considered as highly druggable protein cavity [17].

3.7 Retrieval of Ligands

11,993 drug-like molecules with Tonimoto cutoff level of 60% were retrieved from Zinc database(Sterling and Irwin, 2015). Then partial charges were calculated and energies of these compounds were minimized using energy minimization algorithm with default parameters. All minimized structures were saved in .mdb file. Then these prepared ligands were used as an input file for molecular docking (Wadood et al., 2014).

BioAssay stores biological activity data of the chemical structures. The structures of these ligands were constructed using MOE-Builder tool. Then these compounds were modelled and their partial charges was calculated.

Then the energy of these compounds was minimized using energy minimization algorithm with default parameters. All minimized structures were saved in .mdb

file. Then for docking in MOE-DOCK, these prepared ligands were used as an input file [68].

3.8 Preparation of Protein for Docking

In this study, the protein molecule was obtained from SwissModel. In MOE tool, the solvents were removed and 3D protonation of the drug targets was done in MOE tool. Then the energy of these molecules was minimized using energy minimization algorithm with default parameters. The minimized structures were further used as template for molecular docking.

3.9 Drug Targets Molecular Docking

The drug targets molecular docking was carried out in MOE using MOE-Dock. It predicted the favorable binding possess of selected ligands active sites of drug targets. Default parameters were selected for Molecular docking. After the docking was done, we analyzed the best poses for hydrogen bonding/π–π interactions and then root mean-square deviation (RMSD) was calculated in MOE [68]. For better understanding of these interaction Chimera tool was used which produces 3D representation of these protein-ligand interaction. This tool is freely available non-commercial tool. It works on multiple platforms such as Microsoft windows, Linux, Mac etc. This tool is design for understanding the sequence-structure relationship of protein and allow us for 3D visualization and structure analysis of proteins. It provides better high quality images and it is also able to handle the molecule on all scales such as large molecules assemblies. it can be use through command line or graphical interface which is available at https://www.cgl.ucsf.edu/chimera/ [97].

This software provides better 3D visualization of molecular and related data which includes several magnificent features such as density maps, supramolecular assemblies, molecular dynamics and multiple sequence alignments. The user can create

images in better graphics for publications or presentation. You can download the image in several platforms such as .tiff, pdf or png. In addition to promoting core visualization, the platform is expressly designed for extensibility, allowing the implementation of additional desirable features by outside developers. Current extension such as Multiscale Models which is used ti visualize large scale Molecular assemblies for example viral coat, user can also screen docked ligand orientation by using View Dock, visualize density maps by using Volume Viewer, can display sequence alignments the Multalign Viewer [97].

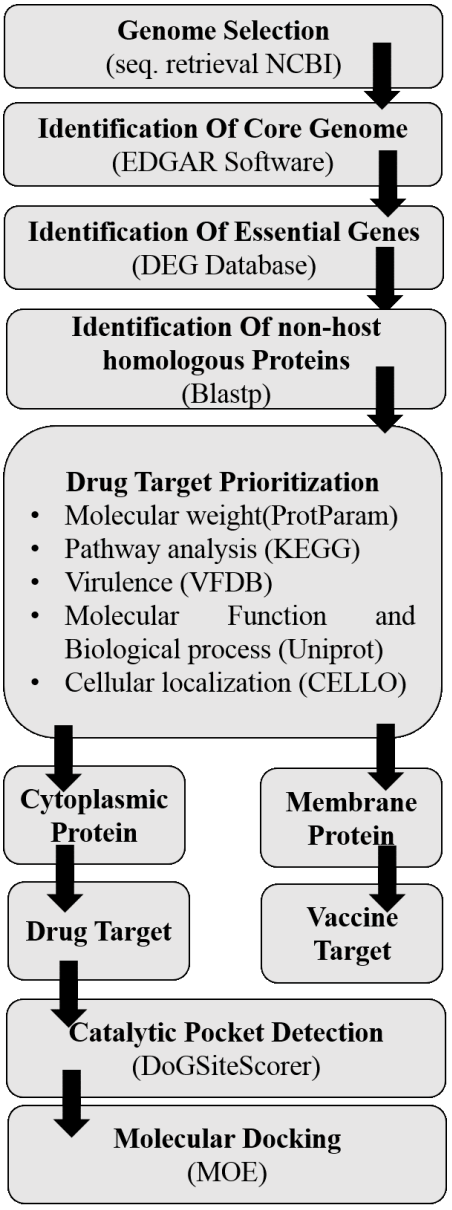

FIGURE 3.1: Methodological steps to identify drug targets in *Sg* using in-silico approach

Chapter 4

Result and Discussion

To prevent the onset of this disease new therapeutics targets were required. The selected complete strains were used to identify core genomes using pan genome analysis. Then the core genomes were subjected for subtractive genomic analysis which was used to identify the essential and non-host homologous genes. These essential non-host homologous genes were subjected for drug prioritization which identified the cytoplasmic and membrane protein with their druggability properties like molecular weight, pathway, functionality, subcellular localization and virulence.

The cytoplasmic proteins were used as drug target while membrane proteins were identified as vaccine targets. These targets were further used for the protein-ligand interactions. This led us to identify novel therapeutic targets. This sections includes the identification of core genome from Sg strains which is further divided into selection of genome and identification of core genome using pan genome, and subtractive genomic analysis.

4.1 Identification of Core Genome of *Sg* Strain

Core genome identification includes selection of the genomes and identification of core genes by sing pan genomic approach.

4.1.1 Selection of Genome

The seven strains of *Sg* were selected for this study. The selection was based on their complete genome to have accuracy in our result. The gene and proteins of these seven strains of *Sg* were retrieved from NCBI (https://www.ncbi.nlm.nih.gov).

4.1.2 Identification of Core Genomes Using Pan-Genomic Approach

The pan genomes were identified using EDGAR tool. UCN34 strain was selected as reference genome and the rest of the strains were compared to the reference strain. The result was in tabular form which includes the locus tag and description of genes along with the fasta file (DNA and Protein). The total genes that it identified in pan genome was 3,242 genes, out of these 1,138 were core genes.

4.2 Subtractive Genomic Analysis from Identified Core Genomes

Subtractive genomic analysis includes two steps identification of non-homologous proteins and identification of essential genes.

4.2.1 Identification of non-host Homologous Proteins

The non-host homologous proteins were identified by NCBI-BLASTp using default parameters against human genome to filter out the non-host homologous proteins. All the protein sequences of 1,138 of core genome were subjected for Blastp, out of which only 1,115 proteins were non-host homologous. Non-homologous proteins are those proteins which are present in pathogen but absent in host (human). For this study this step was very important to avoid any side effect while targeting the drug targets or vaccine targets.

4.2.2 Identification of Essential Genes

The core 1,115 non-hosts homologous proteins were subjected to the Database of Essential Genes (DEG) for the identification of essential protein, through which a final set of 18 proteins, were obtained shown in table 4.1.

TABLE 4.1: List of Pathogen-Essential Non-Homologs Proteins

Query_d	Subject_d	Pct_ dentity	Protein
GALLO_S 00005	DEG10330356	92.857	Chromosomal replication initiator protein DnaA
GALLO_ S00200	DEG10200056	80.769	Glucan-binding protein C
GALLO_ S00610	DEG10010101	54.688	Membrane protein insertase YidC
GALLO_ S00675	DEG10380051	53.659	Transcriptional regulator CtsR
SGGBAA2 069_S00890	DEG10280041	51.448	PTS fructose transporter subunit IIA
GALLO_ S00830	DEG10470198	50	Penicillin-binding protein 2A
SGGBAA2 069_S01250	DEG10180105	47.283	AraC family t ranscriptional regulator
GALLO_ S01215	DEG10110082	45.455	DNA polymerase III subunit alpha
GALLO_ S01760	DEG10060346	44	50S ribosomal protein L28
GALLO_ S01960	DEG10470004	41.793	2-isopropylmalate synthase
GALLO_ S02145	DEG10080178	40.355	Ribosome-binding factor A

GALLO_ S02350	DEG10050423	39.623	Amino acid ABC transporter substrate-binding protein, \ PAAT family /amino acid ABC transporter membrane protein, PAAT family
GALLO_ S02740	DEG10300014	38.71	DNA-binding response regulator
GALLO_ S02995	DEG10430209	38.197	16S rRNA methyltransferase B
GALLO_ S03395	DEG10180247	36.364	Glutamine ABC transporter permease
GALLO_ S03550	DEG10450136	35.789	Penicillin-binding protein 2B
GALLO_ S03570	DEG10460377	35.294	UDP-N-acetylmuramoyl -tripeptide–D-alanyl -D-alanine ligase
GALLO_ S03600	DEG10050249	35.135	1-acyl-sn-glycerol- 3-phosphate acyltransferase

4.3 Drug Prioritization

Determining the potential therapeutic targets there are several factors such as molecular weight, molecular and cellular function, virulence and pathway analysis.

4.3.1 Molecular Weight

The MW of the proteins was calculated by Protparam. It calculates the MW in g/mol when it was converted into the kDa, all proteins MW was less 100kDa. The MW of proteins is shown in table 4.3.

4.3.2 Subcellular Localization

Out of 18 proteins 12 proteins were cytoplasmic, 4 were membrane protein and 2 was extracellular protein respectively. The subcellular localization of protein is shown in table 4.3.

4.3.3 Identification of Virulence of Target Proteins

All target proteins were found to be virulent when those targets were Blast against the VFDB database. The list of virulent target proteins is shown in table 4.3.

4.3.4 Identification of Molecular and Biological Function

In this study the molecular and biological function were retrieved from uniport. All the target proteins molecular and biological function is shown in table 4.2.

4.3.5 Pathway Analysis

In this study KEGG database was used to determine the pathway of the targeted proteins which showed one promising direction for inference of drug targets. The pathway of the targeted protein is shown in table 4.3.

TABLE 4.2: Drug and Vaccine Target Prioritization Parameters and Functional Annotation of 20 Essential Non-Host Homologous Putative Targets.

Uniprot ID	**Protein**	**Gene**	**Biological Function**	**Molecular Function**
A0A139R4E3	Chromosomal replication initiator protein DnaA	dnaA	ATP binding, DNA replication origin binding	DNA replication initiation, regulation of DNA replication

F5X0V5	Glucan-binding protein C	gbpC	-	-
A0A380K0J7	Membrane protein insertase YidC	yidC1	membrane insertase activity	protein transport
F5WXJ0	Transcriptional regulator CtsR	ctsR	DNA binding	regulation of transcription, DNA- templated
A0A3E2SCT8	PTS fructose transporter subunit IIA	DW662_4200	phosphoenolpyruvate-dependent sugar phosphotransferase system	-
A0A380K3P1	Penicillin-binding protein 2A	pbp2A	-	penicillin binding, transferase activity, transferring acyl groups
A0A380K803	AraC family transcriptional regulator	melR	Transcription, Transcription regulation	DNA-binding transcription factor activity, sequence-specific DNA binding

A0A38 0K8Y7	DNA poly erase III subunit alpha	dnaE	DNA replication	3'-5' exonuclease activity, DNA-directed DNA polymerase activity, nucleic acid binding
A0A06 0RG19	50S ribosomal protein L28	rpmB	translation	structural constituent of ribosome
D3HCJ2	2-isopro pylmalate synthase	leuA	leucine biosynthetic process	2-isopro pylmalate synthase activity
F5WZ36	Ribosome-binding factor A	rbfA	maturation of SSU-rRNA	-
A0A139 QNY0	Amino acid ABC tran sporter substrate-binding protein, PAAT family/amino acid ABC transporter membrane protein, PAAT family	SAM N05 6603 28 10839	nitrogen compound transport	transme mbrane transporter activity

A0A139 R8A5	DNA-binding response regulator	DW662 02135	phosph orelay signal transduction system, regulation of transcription, DNA-templated	DNA binding
A0A1S 5WAD9	16S rRNA methyltra nsferase B	BTR42 02745	regulation of transcription, DNA-templated	RNA binding, rRNA methyltra nsferase activity
A0A36 8UI96	Glutamine ABC transporter permease	CAC02 01540	nitrogen compound transport	transmembrane transporter activity
F5WZQ3	Penicillin-binding protein 2B	pbp2B	-	penicillin binding, transferase activity, transferring acyl groups
F5WZQ7	UDP-N-acetylm uramoyl-tripeptide–D-alanyl-D-alanine ligase	murF	cell cycle, cell division, cell wall organization, peptidoglycan biosynthetic process, regulation of cell shape	ATP binding, UDP-N-acetylmu ramoyl-tripeptide -D-alanyl-D-alanine ligase activity

A0A380K5L8	1-acyl-sn-glycerol-3-phosphate acyltransferase	plsC	-	1-acylglycerol-3-phosphate O-acyltransferase activity

TABLE 4.3: Drug and Vaccine Target Prioritization Parameters and Functional Annotation of 20 Essential Non-Host Homologous Putative Targets.

Uniprot ID	**Sub-cellular Localization**	**Virulent**	**Molecular Weight**	**Pathway Analysis**
A0A139R4E3	Cytoplasmic	yes	51401.48	Two component system
F5X0V5	Extra Cellular	yes	47224.62	No Hit
A0A380K0J7	Membrane	yes	84059.88	No Hit
F5WXJ0	Cytoplasmic	yes	7598.78	Transcriptional regulator of stress heat shock response
A0A3E2SCT8	Cytoplasmic	yes	14982.13	No Hit
A0A380K3P1	Cytoplasmic	yes	84763.57	beta-Lactam resistance
A0A380K803	Cytoplasmic	yes	31811.17	No Hit
A0A380K8Y7	Cytoplasmic	yes	165491.77	DNA replication, Mismatch repair, Homologous recombination

A0A060RG19	Cytoplasmic	yes	6883.21	Ribosome
D3HCJ2	Cytoplasmic	yes	33415.6	Biosynthesis of secondary metabolites, 2-Oxocarboxylic acid metabolism, Biosynthesis of amino acids, Valine, leucine and isoleucine biosynthesis, Pyruvate metabolism, Metabolic pathways
F5WZ36	Cytoplasmic	yes	13409.48	No Hit
A0A139QNY0	Membrane	yes	30478.71	No Hit
A0A139R8A5	Cytoplasmic	yes	23939.71	No Hit
A0A1S5WAD9	Cytoplasmic	yes	19761.96	No Hit
A0A368UI96	Membrane	yes	25348.3	No Hit
F5WZQ3	Extra Cellular	yes	77095.67	β-Lactam resistance
F5WZQ7	Cytoplasmic	yes	50278.43	Vancomycin resistance, Peptidoglycan biosynthesis, Metabolic pathway Lysine biosynthesis

A0A38 0K5L8	Membrane	yes	28686.68	Glycerolipid metabolism, Glyceroph ospholipid metabolism, Metabolic pathways, Biosynthesis of secondary metabolites

4.4 Protein-Ligand Interaction

For the protein-ligand interaction, the first step is to detect the catalytic pockets of identified drug target which show the binding sites of these drug targets for the binding to the corresponding ligand and then to perform molecular docking.

4.4.1 Catalytic Pocket Detection

For the catalytic pocket detection DoGSiteScorer is used. For identified drug targets those pockets were selected whose druggability score was greater than 0.6. Druggability score above that 0.60 is considered to be good but score above than 0.8 is favored.

4.4.2 Molecular Docking

For molecular docking of the drug targets, it includes selection of ligands, 3D structure prediction of the targeted protein and protein-ligand docking.

4.4.2.1 Selection of Ligands/ Compounds

11,993 drug-like molecule with Tanimoto cutoff level of 60% was retrieved from zinc database. The structures of these ligands were constructed using MOE-Builder tool. Then these compounds were modelled and their partial charges was calculated.

Then the energy of these compounds was minimized using energy minimization algorithm with default parameters. All minimized structures were saved in .mdb file. Then for docking in MOE-DOCK, these prepared ligands were used as an input file [68].

4.4.2.2 3D Structure Prediction

The structure of all of the targets proteins was predicted as the structure of these proteins were not available in protein databank (PDB). SwissModel web tool was used to predict the 3D of these targeted proteins [69]. The workflow of this tool includes the main steps first is data input.

In data input all targeted proteins sequences (Fasta Format) were provided to this tool. Then second step is Data Search, in this for the provided data it searches its evolutionary related protein structure against Swiss-Model Template Library (SMTL). It used two databases while performing this task first is BLAST which is fast and sufficiently accurate for closely related templates and second is HHblits which adds sensitivity to the remote homology structures[69].

Remote homologs are those pair of proteins which have same structure and functions but lack easily detectable sequence similarity[70]. After template search then comes third step which is template selection. It provides us with the all top ranked templates whose quality was estimated by Global Model Quality Estimate (GMQE). The templates were selected whose sequence similarity score was high. Then the fourth step is model building upon the selected template it builds the 3D structure for targeted proteins.

4.4.2.3 Validation of 3D Structures

All the 3D structures quality was further validated by using RAMPAGE and ERRAT tool shown in table 4.4. Rampage stands for RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression. This tool does Ramachandran plot analysis and giving us the validity score for each target protein 3D structure. The score greater than 80 is considered good [71]. In this tool we provided it .pdb file of the 3D structure of the targeted protein, score for all the targeted protein was greater than 80. For the more accurate data the second tool for validation was ERRAT tool. It is also an online structure evaluation tool. The quality factor of the 3D structure above then 37% is considered good[60]. The .pdb file was provided to this tools and quality factor for all the predicted 3D structure were greater than 80.

Table 4.4: Validation Score of Models from Rampage and ERRAT

Sno.	Protein Name	Errat	Rampage
1.	16S rRNA methyltransferase B	90.6699	92.30%
2.	PTS fructose transporter subunit IIA	88.0435	90.80%
3.	50S ribosomal protein L28	74.0741	87.50%
4.	Chromosomal replication initiator protein DnaA	93.6747	92.60%
5.	Penicillin-binding protein 2A	93.6823	91.30%
6.	DNA polymerase III subunit alpha	89.1	88.90%
7.	AraC family transcriptional regulator	100	97.00%
8.	DNA-binding response regulator	93.0693	92.00%
9.	Transcriptional regulator CtsR	100	100.00%
10.	Ribosome-binding factor A	100	96.90%
11.	UDP-N-acetylmuramoyl-tripeptide –D-alanyl-D-alanine ligase	94.7248	94.20%
12.	2-isopropylmalate synthase	92.766	94.90%

4.4.2.4 Docking

Docking was performed against 12 drug targets with 11,993 Zinc drug-like compound in MOE tool. Out of which top 100 molecules were selected. Then these top 100 molecules were redock from which top 10 molecules was selected. For each protein 1 best interaction was selected from these top 10 molecules. The best interaction of each protein-ligand is drawn in chimera. The analysis and biological significance of each of the predicted protein-ligand interaction are described here: **16S rRNA methyltransferase B (BTR42-02745)** is a protein which play an important role in methylation of cytosine at position 967 of 16S rRNA. The structure of this protein consist of active sites in which two conserved cysteine residues are present. These cysteine residues are located near the activated methyl of cofactor. One of the cysteine residues act as a catalytic nucleophile and other play an important role in methyl transferase mechanism (Foster et al., 2003). The top 10 best confirmations are shown in table 4.4 along with their ZincID, number of interactions, Interacting Residues and minimized energy. The residues Lys 285, Lys 339 and Cys330 were found to interact with active ligand (ZINC01532584). The interaction of 16S rRNA methyltransferase B with ZINC01532584 is shown in figure 4.1.

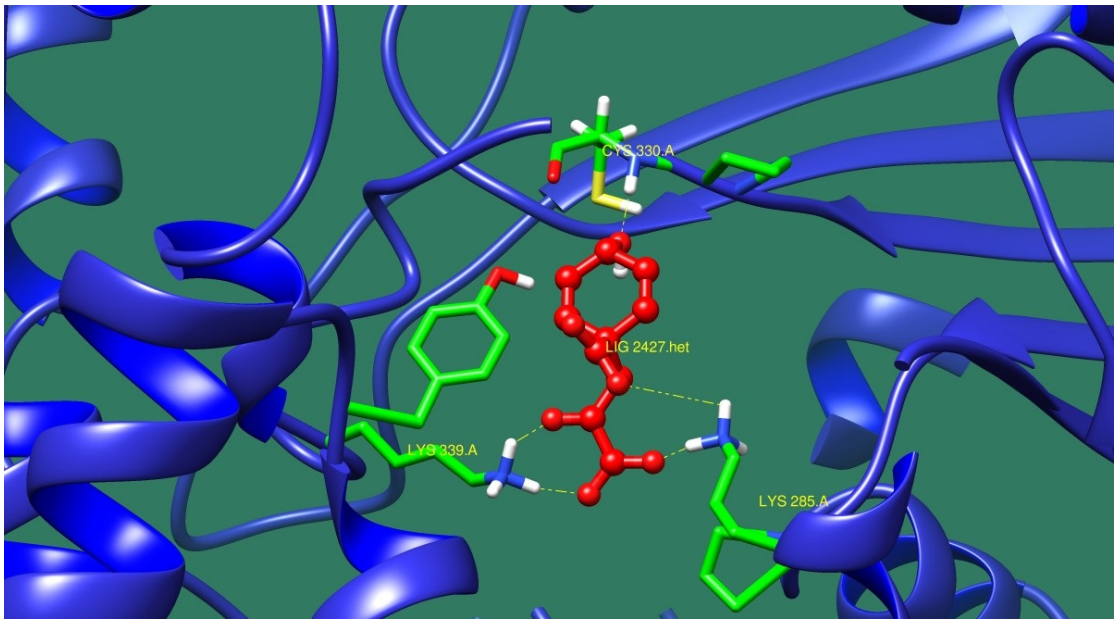

FIGURE 4.1: Interaction OF 16S rRNA methyltransferase B with ZINC 01532584

TABLE 4.5: ZincID, Minimized energy, Scientific names of Compounds, Number of interaction and Interactive Residues for 16S rRNA methyltransferase B

Zinc ID	Scientific names of Compounds	Number of Interactions and Interacting \ Residues	Minimized Energy
ZINC 05835424	1,3-Cyclohexadiene-1 -carboxylicacid, 6-amino-5- hydroxy-(9CI)	Ser 238,Asp327, Lys 263, Ala 328	-12.453
ZINC 13650894	LY 233053	Lys 339, Lys 285, Lys 263, Asp 327	-14.373
ZINC 13520246	Dimethyl acid pyrophosphate	Lys 263, Gly 262, Ser 238	-14.238
ZINC 07001187	2-(7,8-dimethyl -5-oxo-9-thia-2,4 -diazabicyclo[4.3.0] nona-2,7,10-trien -3-yl)-3-(2-furyl) prop-2-enenit	Arg 338, Asp 235, Lys 263	-18.2
ZINC 32714665	3-[4-(dimethy laminomethyl eneamino)-1,2,4 -triazol-1-ium-1- yl]propane-1-sulfonic	Ala 328, Lys 263, Asp 327, Gly 262	-32.289
ZINC 1404930	Not known	Tyr 382, Lys 285, Lys 339	-14.545
ZINC 01711849	[1,2,4,5]Tetrazine- 3,6-dicarboxylic acid disodium salt	2/ Lys 346, 2/Lys 285	-0.952

TABLE 4.5: ZincID, Minimized energy, Scientific names of Compounds, Number of interaction and Interactive Residues for 16S rRNA methyltransferase B

Zinc ID	Scientific names of Compounds	Number of Interactions and Interacting \ Residues	Minimized Energy
ZINC 01532584	Prephenate	2/ Lys 285, 2/ Lys 339, Cys 330	-22.145
ZINC 05181663	(hydroxy-methox y-BLAHyl)urea	Ser 331, Lys 339, Lys 285	-21.977
ZINC 44551376	N-[(Z)-[(4R)-4-methylimidazolidin -2-ylidene]amino] pyrazine-2-carbox amide	2/ Asp 341, Lys 339, Ser 27, Asn 28	-8.625

Chromosomal replication initiator protein DnaA (dnaA) is a protein which play an important role in initiation and regulation of chromosomal replication. In DNA regulation the initiation process is the key event in the cell cycle in all organism. The initiation of replication starts at the site of origin which is recognized and processed by the initiator protein.

The structure of this protein consist of nucleotide binding folds with the long helical connector to all-helical DNA binding domain. The conserved motif of this protein provide information about two most important steps in origin processing which are binding of DNA and homo-oligomerization (The structure of bacterial DnaA: implications for general mechanisms underlying DNA replication initiation). Table 4.6 presents top 10 protein-ligand interaction with ZincID, Minimized energy, Number of interactions and interactive residues. ZINC71782058 was predicted as most active lead compound against Chromosomal replication initiator protein DnaA (dnaA). The protein-ligand interaction is shown in figure 4.2.

TABLE 4.6: ZincID, Minimized energy, Scientific names of Compounds, Number of interactions and Interactive Residues for Chromosomal replication initiator protein DnaA

Zinc ID	Scientific names of Compounds	Number of interaction & Interacting Residues	Minimized Energy
ZINC 05839384	2,2-diethoxy-4-methyl-2-methy ldisulfanyl-1-oxa-3-thia-2$l^{5}-\ phosphacyclop entan-5-one	Lys 291, Asn 120, Lys 115	-12.715
ZINC 07089629	MFCD02956395	Lys 291, Asn 120	-15.34
ZINC 13540203	Creatine phosphate	2/ Arg 417, Lys 412, Asp 312	-21.772
ZINC 71618824	Not known	2/ Arg 417	-14.766
ZINC 71782058	Not known	4/ Arg 41,Lys 412	-24.383
ZINC 72281564	N-(2-ethylbutyls ulfonyl)-2-fluor o-pyridine-4-car boxamide	2/ Lys 291, Asn 120	-16.347
ZINC 01585185	2,2,3,3-tetramet hoxyindeno[1,2-b][1,4]dioxin-9-one	3/ Lys 291, Asn 120, Lys 115	-12.005
ZINC 01152242	5-(2-pyridylsu lfamoyl)-2-furoate	Tyr 116, Lys 291	-13.191
ZINC 00387687	MFCD04154099	Lys 115, Glu 294	-0.384

TABLE 4.6: ZincID, Minimized energy, Scientific names of Compounds, Number of interactions and Interactive Residues for Chromosomal replication initiator protein DnaA

Zinc ID	Scientific names of Compounds	Number of interaction & Interacting Residues	Minimized Energy
ZINC 01844424	(6R)-2-[(Z)-3H-1,3-benzothiazol-2-ylideneamino]-4-keto-5,6-dihydro-1H-pyrimidine-6-carboxylate	Lys 291, Asn 113	-22.083

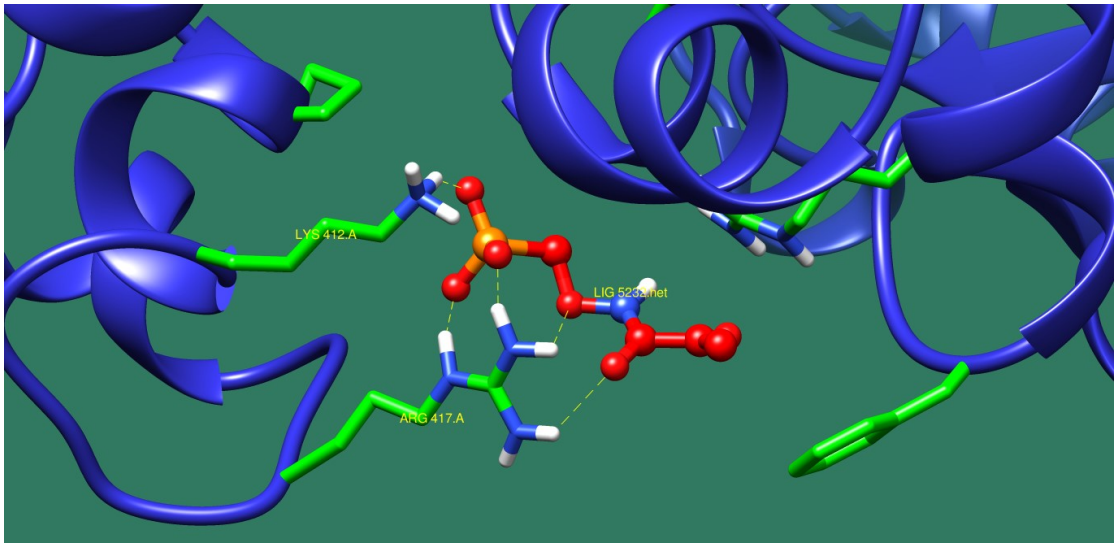

FIGURE 4.2: Interaction of Chromosomal replication initiator protein DnaA with ZINC71782058

Transcriptional regulator CtsR (ctsR) is an important repressor which regulates the transcription of class III stress genes in Gram-positive bacteria. CtsR controls the expression of genes encoding for chaperons and proteases. These genes play an important role in protein quality control system of bacteria. The structure of this protein consist of N-terminal DNA binding domain and C-terminal dimerization domain. N-terminal DNA binding domain consist of HTH folds and

C-terminal dimerization domain consist of α-helices organized in four helix bundle. This protein also play an important role pathogenicity as it provides benefit to the bacteria during its stress condition and improves the survival chances for bacteria (Fuhrmann et al., 2009). Top 10 lead molecules against this protein are shown in table 4.7 consisting ZincID, minimized energy, numbers of interactions and interacting residues. The best interaction was shown by ZINC79090716 as shown in figure 4.3.

TABLE 4.7: ZincID, Minimized energy, Scientific names of Compounds, Number of interactions and Interactive Residues for Transcriptional regulator CtsR

Zinc ID	Scientific names of Compounds	Number of interaction and Interacting Residues	Minimized Energy
ZINC 05839384	2,2-diethoxy-4-methyl-2-methyldisulfanyl-1 -oxa-3-thia-2$l^{5}- phosphacyclopentan-5-one	3/ Asp 124	-22.285
ZINC 06962237	4,6-dimethyl-N-[(3-pyrrolidin-1-yl-1- cyclohex-2-enylidene) amino]pyrimidin-2-amine	Thr 111	-19.713
ZINC 19510011	6-[(2,6-dioxo3H-pyrimidine-4-carbonyl)amino]hexanoic	2/ Arg 113	-9.787
ZINC 71603173	Not known	Glu 114	-79.985
ZINC 77504434	(6S)-6-[[2-(5-methyl-2-furyl) ethylamino]methyl] -1,4-oxazepan-6-o	Thr 111	-16.607
ZINC 79090716	1-[2-(benzenesulfonyl)ethyl] -4,5-dimethyl-imidazole	Thr A111, Thr B111, Glu 114	-32.201

TABLE 4.7: ZincID, Minimized energy, Scientific names of Compounds, Number of interactions and Interactive Residues for Transcriptional regulator CtsR

Zinc ID	Scientific names of Compounds	Number of interaction and Interacting Residues	Minimized Energy
ZINC 01672834	4-methylbenzene sulfonic-acid-(4-amino-1-pyrindan -1-ium-2-yl)-ester	Thr B111	-20.743
ZINC 04352554	MFCD00014510	Thr B111	-6.881
ZINC 655337127	4,6-Difluoro-1H-benzo [d]imidazol-2(3H)-one	Thr A111, Thr B111	-10.149
ZINC 65337127	Not known	Thr A111	-17.328

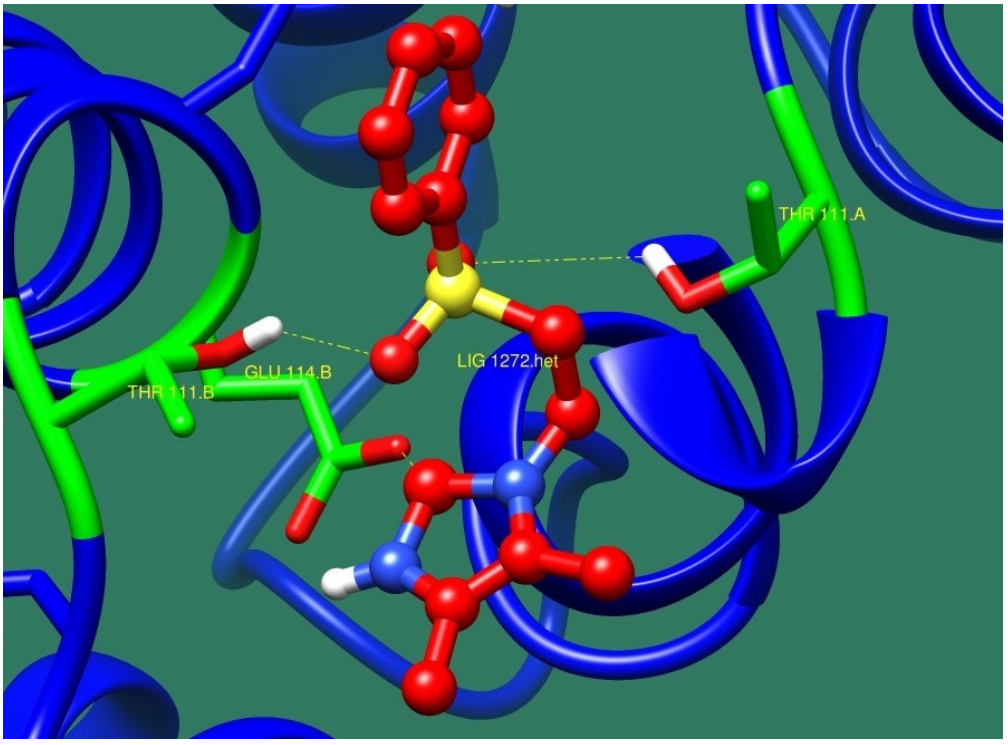

FIGURE 4.3: Interaction of Transcriptional regulator CtsR with ZINC79090716

PTS fructose transporter subunit IIA (DW662-04200) is protein which is involved in phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS). In bacteria it is a major carbohydrate transport system. PTS catalyzes the translocation with naturally occurring phenomenon of phosphorylation of sugar and hexitols and it also regulates the metabolism in response to the availability of carbohydrates.

It consists of two protein HPr and enzyme I protein. These are the cytoplasmic protein; in which first enzyme I transfers phosphoryl groups from phosphoenolpyruvate to phosphoryl carries protein HPr. Then this HPr further transfers the phosphoryl group to different transport complexes. PTS fructose transporter subunit IIA belongs to the fructose-Manitol family. This is large and complex family which consist of several sequenced fructose and mannitol-specific permeases and putative permeases of unknown specificities.

This family have three domain IIA, IIB and IIC from which the most specific domain is IIA for the fructose PTS transporters (Siebold et al., 2001). The top 10 protein-ligand interaction is shown in table 4.8 and the best interaction is shown in figure 4.4 with ZINC01638334.

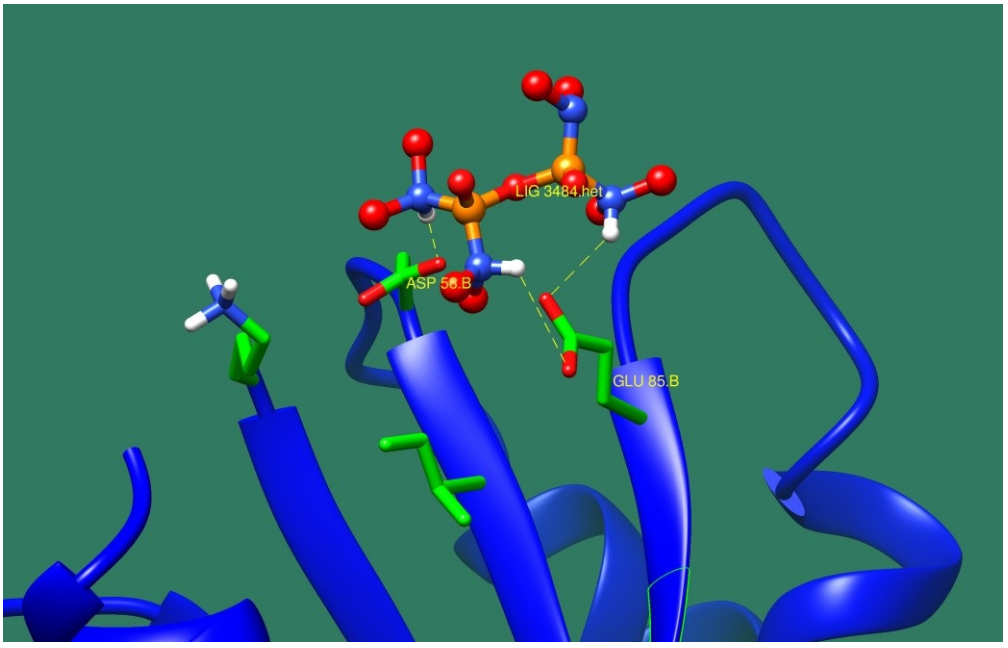

FIGURE 4.4: Interaction of PTS fructose transporter subunit IIA with ZINC01638334

TABLE 4.8: ZincID, Minimized energy, Scientific names of Compounds, Number of interactions and Interactive Residues for PTS fructose transporter subunit IIA

Zinc ID	Scientific names of Compounds	Number of interactions and Interacting Residues	Minimized Energy
ZINC18033182	N'-dimethoxy phosphoryl-N ,N-dimethyl- formamidine	3/ Asp 58 , Glu 85	-52.033
ZINC32714665	3-[4-(dimethylamino methyleneamino)-1,2, 4-triazol-1-ium-1-yl] propane-1-sulfonic	His 83, Glu 85, Asp 58	-65.244
ZINC17004087	(6aR)-2,3,4,6,6a, 8-hexahydro-1H -quinolino[3,4-c] cinnolin-7-one	Glu 85, Tyr 87, Asp 58	-12.11
ZINC72145573	(3aS,6aR)-2-(5,6,7,8- tetrahydro-4H-pyrazolo [1,5-a][1,4]diaze pin-2-ylmethyl) -1,3,3a,6a-tetrahydropyrr	2/ Lys 3, Glu 85, Asp 58	-19.982
ZINC71780811	Not Known	2/ Lys 118 , Gln 28, Glu 22	-23.974
ZINC01638334	Not Known	Asp 58, 2/ Glu 85	-27.139
ZINC01613419	dimethyl-[2,3,4-tris (dimethylamino) tetraphosphetan -1-yl]amine	2/Glu85, 2/Asp58	-22.839

TABLE 4.8: ZincID, Minimized energy, Scientific names of Compounds, Number of interactions and Interactive Residues for PTS fructose transporter subunit IIA

Zinc ID	Scientific names of Compounds	Number of interactions and Interacting Residues	Minimized Energy
ZINC04261883	BOP	3/ Glu 85, His 83	-11.661
ZINC38292458	2-aminoethyl sulfonylazide	2/ Glu 85, Asp 58	-15.396
ZINC49625635	2,3-dihydroi midazo[1,2-c] isoxazolo[4,5-e] pyrimidine-9-carboxylic acid	2/ Asp 58,2/Glu85	-51.252

Penicillin-binding protein 2A (pbp2A) is a transpeptidase which catalyzes the cell wall crosslinking in the face of challenge by β-Lactam antibiotics. This protein activation is regulated by active site at which the cross linking take place (Fishovitz et al., 2014). Through pathway analysis it is clear that it is involved in β-lactam resistance pathway. β-lactam antibiotic is the most used group of antibiotics, which exert its effect by interfering with the bacterial cell wall by structural cross linking of peptidoglycan. This protein has already been reported as β-lactam resistant.

This antibiotic resistance is due to the inactivation of the enzymes, change β-lactam targets of pbp, change in porins the transport of β-lactam to periplasmic is reduced and use of efflux pump for the exclusion of β-lactam (Kocaoglu and Carlson, 2015). The top ranked lead compounds are given in table 4.9 where compound ZINC16942644 was predicted as best on basis of minimized energy and number of interactions made (figure 4.5).

TABLE 4.9: ZincID, Minimized energy, Scientific names of Compounds, Number of interactions and Interactive Residues for Penicillin-binding protein 2A

Zinc ID	Scientific names of Compounds	Number of interaction and Interacting Residues	Minimized Energy
ZINC05567030	N,N'-dinitroso-1H-1,2, 4-triazole-3,5-diamine	2/ Asp408	-13.81
ZINC22048956	(2R)-2-amino-3-[(1-hydroxy-2,2,5,5-tetramethyl-pyrrol-3-yl)methylsulfanyl]propanoic	Tyr 456, Glu 421, Gln 424	-15.356
ZINC19799513	2-(1,3-dimethyl-2-oxo-benzimidazol-5-yl)-3H-quinazolin-4-one	Lys 166, Asp 382	-19.047
ZINC17004087	(6aR)-2,3,4,6,6a,8-hexahydro-1H-quinolino [3,4-c]cinnolin-7-one	2/ Asp 382, Glu 381	-16.385
ZINC18045201	1,3- diphosphinane- 1, 3- diol1,3- dioxide	2/ Arg 443, Gln 424	-16.838
ZINC20502353	[2-(2-methylpyridin-1-ium-1-yl)-1-phosphono-ethyl] phosphonic	Tyr 456,2/ Gln 424	1.255
ZINC20070370	3-(3,5,6-trimethylthieno[2, 3-d]isothiazol-2-ium-2-yl) propane-1-sulfonic	Gly 425, Ser 424	-6.277
ZINC32628102	5-(3-aminophenyl)-8, 9-dihydro-2 H-pyridazino[4, 5-a]pyrrolizine-1,4(3H,7H)-dione	Arg 443, Gly 425	-13.827

TABLE 4.9: ZincID, Minimized energy, Scientific names of Compounds, Number of interactions and Interactive Residues for Penicillin-binding protein 2A

Zinc ID	Scientific names of Compounds	Number of interaction and Interacting Residues	Minimized Energy
ZINC16942644	(2S,3S,3aR,8aR)-5-ethoxy-3-hydroxy-2-(hydroxymethyl) -7-oxo-3,3a,8,8a-tetrahydro-2H-furo[3,2-b][1,4]o	Gln 424, 2/ Gly 425, Ala 423	-3.839

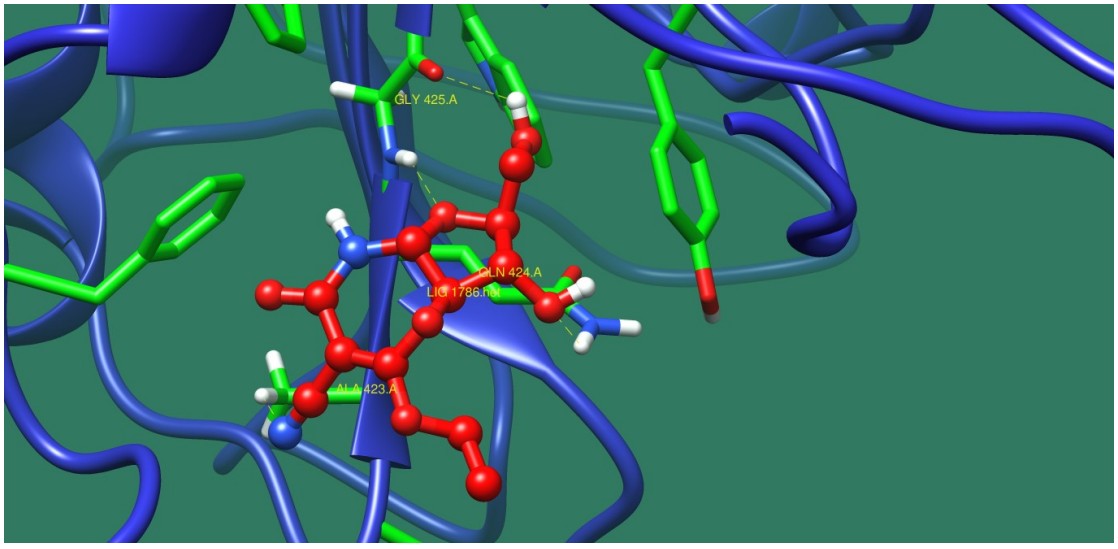

FIGURE 4.5: Interaction of Penicillin-binding protein 2A with ZINC16942644

UDP-N-acetylmuramoyl-tripeptide-D-alanyl-D-alanine ligase (murF) is a protein involved in the biosynthesis of peptidoglycan. Peptidoglycan is the important component of bacterial cell wall and enzymes involved in its synthesis could represent as potential drug target. MurF catalyzes the final step in the biosynthesis of the peptidoglycan in which it adds the D-Ala-D-Ala to the nucleotide precursor UDP-MurNAc-L-Ala-γ-D-Glu-meso-DAP (Hrast et al., 2013). The protein-ligand interaction of the top 10 molecules is shown in table 4.10 and

among these molecules the best interaction was with ZINC14681317 as shown in figure 4.6.

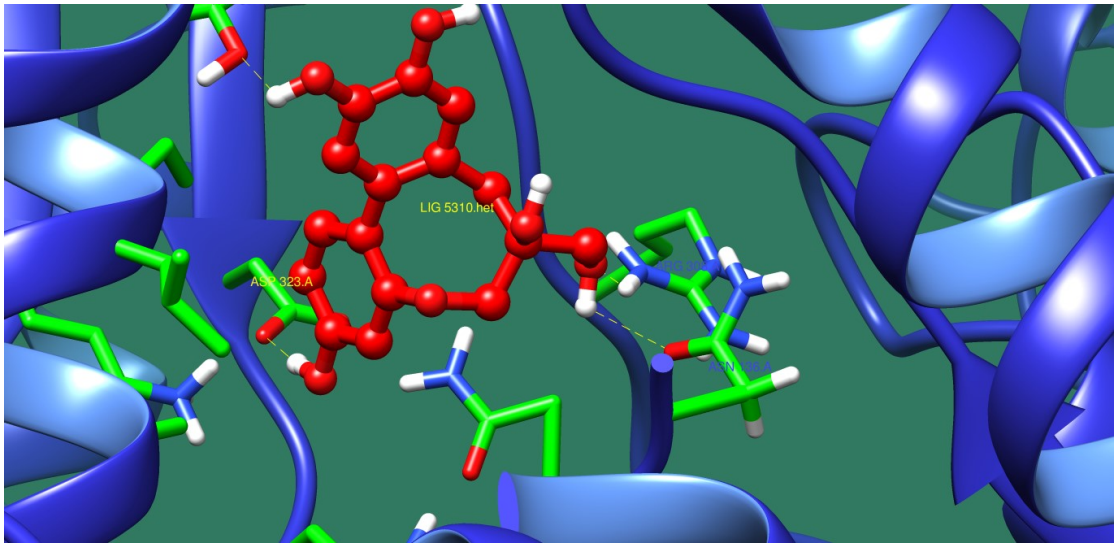

FIGURE 4.6: Intnteraction of UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D-alanine ligase with ZINC14681317

TABLE 4.10: ZincID, Minimized energy, Scientific names of Compounds, Number of interactions and Interactive Residues for UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D-alanine ligase

Zinc ID	Scientific names of Compounds	Number of interaction and Interacting Residues	Minimized Energy
ZINC14681317	Protosappanin B	Thr338, Asp162, Asp323, Asn36, Arg308	-29.34
ZINC05842784	6-[3-methyl-5-(trifluoromethyl)-4,5-dihydropyrazol-1-yl]-4,5,7,8,10-pentazabicyclo[5.3.0]deca-3,5,8,	Asn 137, Asp 162, Glu 138,	-17.576
ZINC05811451	N,N-trimethyl-dioxo-BLAHamine	Thr 309, Asn 134, Arg 308	-12.09

TABLE 4.10: ZincID, Minimized energy, Scientific names of Compounds, Number of interactions and Interactive Residues for UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D-alanine ligase

Zinc ID	Scientific names of Compounds	Number of interaction and Interacting Residues	Minimized Energy
ZINC32714665	3-[4-(dimethylaminomethyleneamino)-1,2,4-triazol-1-ium-1-yl] propane-1-sulfonic	2/ Asn 162, Asn 137	-55.904
ZINC15768374	3-hydroxy-4-[(3-methoxyphenyl) amino]cyclobut-3-ene-1,2-dione	3/ Asp 162	-14.941
ZINC71607274	methyl	Glu138,Asp162,Asn137	-5.876
ZINC77323423	2-methyl-3-(1-methylpyrrolo [2,3-b]pyridin-4-yl) -6H-pyrazolo[1,5-d] [1,2,4]triazin-7-one	Arg 308, Asn 134, Thr 309, Thr 338	-7.009
ZINC73825281	4-amino-2-methyl-5-(4-pyrrolidin-1-yl-2-pyridyl)-1,2,4-triazole-3-thione	2/ Asp 162, 2/ Thr 338	-2.389
ZINC70503687	(3aS,4R,6R,7aR)-4-(4-methoxyphenyl)-3a,4,5,6,7,7a-hexahydro-1H -imidazo[4,5-c] pyridine-6-carboxylic	Thr 309, Thr 338	-41.806

AraC family transcriptional regulator (melR) protein belongs to Arac/XylS Family. This family is transcription regulators and is widely distributed in bacteria. This protein regulates the transcription of several genes and operons which are involved in arabinose catabolism and transport. This protein co-regulates with another transcription regulator which is also involved in degradation of I-arabinose. By binding together these regulators activate the transcription of 5 operons which are involved in transport, catabolism and autoregulation of I-arabinose. Its structure composed of C-terminal DNA binding domain and N-terminal domain. C-terminal DNA binding domain consist of two HTH which is connected with α-Helix and N-terminal domain is responsible for dimerization and binding of I-arabinose. The structure of this reveal that the N-terminal of this protein plays an important role in regulation of arabinose (Rodgers and Schleif, 2009; Fernandez-López et al., 2015; Malaga et al., 2016). Table 4.11 presents the best results against AraC family transcriptional regulator (melR) where ZINC71781167 was predicted as top lead compound as shown in figure 4.7.

TABLE 4.11: ZincID, Minimized energy, Scientific names of Compounds , Number of interactions and Interactive Residues for AraC family transcriptional regulator

Zinc ID	Scientific names of Compounds	Number of interactions & Interacting Residues	Minimized Energy
ZINC06691773	1,2,3,4-tetrachloro cyclobut-2-ene-1, 4-dicarboxylic	3/ Arg 242	1.59
ZINC13552228	Tetramethylaz odicarboxamide	3/ Asp 248	-4.316
ZINC08627906	2,4-dimethyl-2,3,6,7-tetrazabicyclo[3.3.0] octa-3,9-dien-8-imine	Asn 205, 2/ Ile 198	-10.822

TABLE 4.11: ZincID, Minimized energy, Scientific names of Compounds , Number of interactions and Interactive Residues for AraC family transcriptional regulator

Zinc ID	Scientific names of Compounds	Number of interactions & Interacting Residues	Minimized Energy
ZINC15768388	3-(2-furylmethylamino)-4-hydroxy-cyclobut-3-ene-1,2-dione	Lys 245, Asn 205	-3.968
ZINC18164213	(1E)-3-nitro-1-phenylimino guanidine	Tyr 202, Lys 245	-15.968
ZINC18141362	Not Known	Asn 199, Val 241	-12.472
ZINC71603518	Not Known	2/ Asp 248	-6.232
ZINC71781167	Not Known	2/Arg 242, Asn 271	-10.953
ZINC70632388	3-hydroxy-4-pyridin -1-yl-pyrrole- 2,5-dione	Arg 242, Gly 265, Asn 267	-10.747
ZINC71618824	Not Known	2/ Arg 242	-16.738

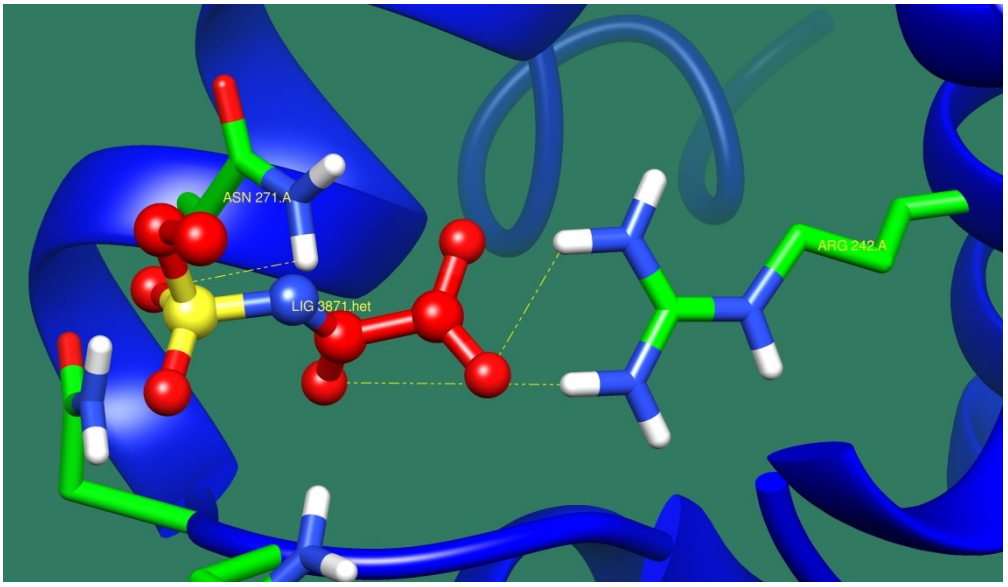

FIGURE 4.7: Interaction of AraC family transcriptional regulator with ZINC71781167

DNA polymerase III subunit alpha (dnaE) is responsible for the replication in bacterial genome. This protein function as tripartite assembly consisting two core polymerases. In *E. coli*, the core polymerases contain the catalytic α-subunit also known as PolIIIα, the 3'-5' exonuclease ε-subunit and the θ subunit whose function is essentially unknown (Wing et al., 2008). From function and pathway analysis, this protein is involved in DNA replication, mismatch repair pathway and homologous recombination. It is located in cytoplasm which mean it could act as drug target. The top 10 interaction of this protein with ligands is shown in table 4.12 along with their ZincID, minimized energy, number of interactions and interactive residues. The binding pocket residues Arg955, Lys553, Gln556 and Arg554 were predicted to contribute in the interaction with lead molecule ZINC38653615 as shown in figure 4.8.

TABLE 4.12: ZincID, Minimized energy, Scientific names of Compounds , Number of interactions and Interactive Residues for DNA Polymerase III subunit alpha

Zinc ID	Scientific names of Compounds	Number of interaction and Interacting Residues	Minimized Energy
ZINC06566417	3-(5-amino-3-methyl -isoxazol-4-yl)-3-hydroxy-5-nitro-indolin-2-one	Arg 955, Arg 554, 2/ Gln 556	-17.247
ZINC08616471	diethyl	Asn 953, Arg 554, Gln 556, Lys 553	-15.181
ZINC05766473	4-(3-hydroxy-1-methyl-2-oxo-indolin-3-yl) benzo[1,3] dioxole-5,6-dione	2/ Arg 955, Arg 554 Gln 556	-22.601

TABLE 4.12: ZincID, Minimized energy, Scientific names of Compounds , Number of interactions and Interactive Residues for DNA Polymerase III subunit alpha

Zinc ID	Scientific names of Compounds	Number of interaction and Interacting Residues	Minimized Energy
ZINC32599342	4-[(5-hydroxy-2-pyridyl)azo] benzenesulfonic	Asn 550, Leu 956, Lys 553, Gly 535	-22.078
ZINC16248201	N-(2-hydroxyethyl) -N-(hydroxy-methyl -phosphoryl)-methyl -phosphonamidic	Gly 535, Leu 956, 2/ Lys 553	-20.68
ZINC00351016	dihydroxyBLA Htrione	Asn 954, Gln 556, Asn 953, Arg 955	-17.012
ZINC00440425	5,5-bis(hydroxymethyl) -3-(p-tolylaminoimino) morpholin-2-one	Lys 553, Arg 554, 2/ Asn 953	-21.827
ZINC05204676	2-[4-(1H-benzoimidazol -2-ylamino)- 6-hydroxy- 1,3,5-triazin-2-yl] acetonitrile	3/ Lyss 538, Lys 919	-14.05
ZINC38653615	N-[(E)-[(3R,3aR,5R, 6S,6aS)-3,5,6-trihyd roxy-3a,5,6, 6a-tetrah ydro-3H-furo[3,2-b] furan-2-ylidene] amino	2/ Arg 955, 2/ Lys 553, Gln 556, 2/ Arg 554	-21.303

TABLE 4.12: ZincID, Minimized energy, Scientific names of Compounds , Number of interactions and Interactive Residues for DNA Polymerase III subunit alpha

Zinc ID	Scientific names of Compounds	Number of interaction and Interacting Residues	Minimized Energy
ZINC44123372	1-cycloheptyl-3-morpholinosulfonyl -urea	2/ Arg 955, Asn 953	-10.264

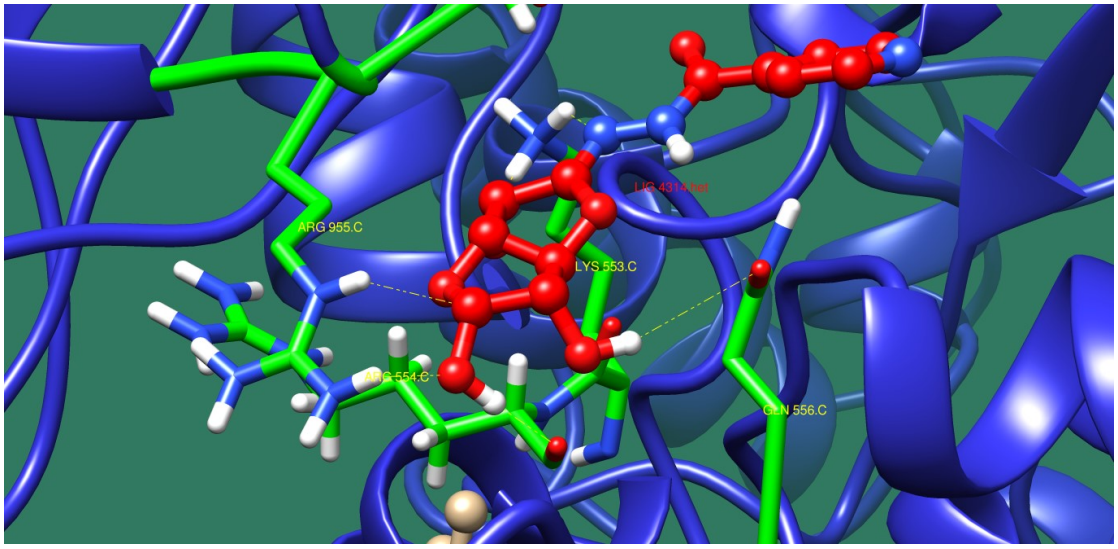

FIGURE 4.8: Interaction of DNA polymerase III subunit alpha with ZINC38653615

50S ribosomal protein L28 (rpmB) protein which plays an important role in the assembly of ribosome. This protein is encoded by rpmB operon. This protein could act as potential drug target as its role in ribosome assembly and it functioning (Aseev et al., 2016). The function analysis also shows its role in translation and structural constituent in ribosomes which makes it a good pharmaceutical target. The top 10 result of 50S ribosomal protein L28 protein is shown in table 4.13 along with their ZincID, minimized energy, number of interactions and interactive residues and the best interaction was observed with ZINC03872713 shown in figure 4.9.

TABLE 4.13: ZincID, Minimized energy , Scientific names of Compounds , Number of Interactions and Interactive Residues for 50S ribosomal protein L28

Zinc ID	Scientific Name of Compounds	Number of Interaction and Interacting Residues	Minimized Energy
ZINC06691773	1,2,3,4-tetrachlorocyclo but-2-ene-1,4 -dicarboxylic	2/Lys 11	1.005
ZINC13540203	Creatinephosphate	2/ Lys 11, 2/Lys30	-35.92
ZINC70632524	3-hydroxy-4-[(2-hydroxy -3,4-dioxo-cyclobuten -1-yl)amino]cyclobut- 3-ene-1,2-dione	Ser 14, Lys 30	-6.21
ZINC77312688	4H-pyran-2,6 -dicarboxylic	Lys 11, Lys 30, Ser 14	-20.899
ZINC78442030	N-methylsulfonyl-2- (2-oxoazocan-1-yl) acetamide	Trp 48, Ala 2	-7.335
ZINC01711849	[1,2,4,5]Tetrazine-3 ,6-dicarboxylic acid disodium salt	2/Lys 11,2/ Lys30	-3.677
ZINC05372521	1,1-dioxo-1,2,5- thiadiazole-3,4- dicarboxylic	2/ Lys 11, Ser 14	-4.936
ZINC03872713	Glyphosate	2/ Lys 11, Ser 14, Thr 12, Lys 30	-17.983
ZINC00053149	3-hydroxy-1,4,6,8- tetrahydropyrazolo [3,4-f]indazol-7-olate	Lys 30, Ala 2	-5.882

TABLE 4.13: ZincID, Minimized energy , Scientific names of Compounds , Number of Interactions and Interactive Residues for 50S ribosomal protein L28

Zinc ID	Scientific Name of Compounds	Number of Interaction and Interacting Residues	Minimized Energy
ZINC03861035	Sodium 3,4,5,6-tetraoxocyclohex-1-ene-1,2-bis(olate)	2/ Ala 2	-8.094

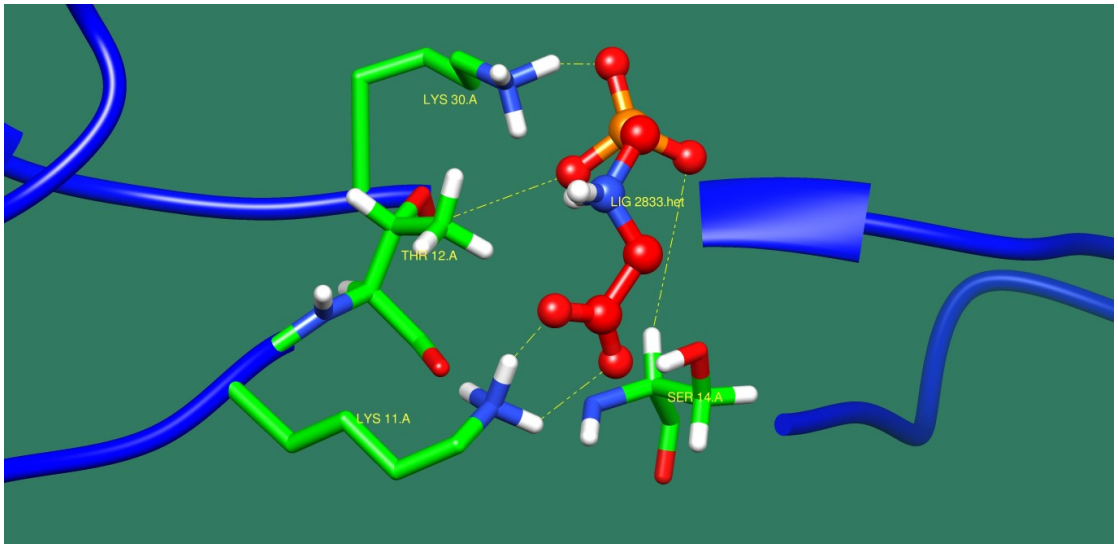

FIGURE 4.9: Interaction of 50S ribosomal protein L28 with ZINC03872713

2-isopropylmalate synthase (leuA) protein catalyzes to form 2-isopropylmalate by the condensation of acetyl group of Acetyl-CoA with 2-oxoisovalerate. It is also involved in biosynthesis of leucine, by synthesizing L-leucine from 3-methly-2-oxobutanate (De Carvalho and Blanchard, 2006). In Mycobacterium tuberculosis biosynthesis of leucine plays an essential role, which is important for the growth of bacteria and so it could act as a potential drug target. The structure of this protein consist two domain N-terminal and C-terminal. N-terminal consist of TIM barrel catalytic domain and C-terminal is a regulatory domain (Koon et al., 2004). The top 10 ligands against 2-isopropylmalate synthase (leuA) protein are shown in table 4.14 along with ZincID, minimized energy, number of interactions and

interactive residue and the best interacting protein-ligand confirmation is shown in figure 4.10.

TABLE 4.14: ZincID, Minimized energy, Scientific names of Compounds , Number of interactions and Interactive Residues for 2-isopropylmalate synthase

Zinc ID	Scientific names of Compounds	Number of interaction & Interacting Residues	Minimized Energy
ZINC08939819	N-(3-nitro-6-oxo-1, 6-dihydropyridin-2-yl)acetamide	Asp 401, Lys 425, Asp 482	-11.066
ZINC05688692	Formamidine, N'-(2-bromoallyl)-N,N-dimethyl-; Formamidine, N,N-dimethyl-N'-(2-bromoallyl)-; LS-69581; N,N-Dimethyl-N'-(2-bromoallyl) formamidine	Asp 401, 2/ Asp 482	-33.487
ZINC32714665	3-[4-(dimethylaminomethyleneamino)-1,2,4-triazol-1-ium-1-yl]propane-1-sulfonic	Lys 487, Asp482, 2/ Asp 401	-41.59
ZINC22056810	1-(2-azidoethoxy)-2-azidoethane	Asp 401, Asp 482, Ala 400, Asp 402	-2.993
ZINC83324781	(E)-4-(1H-imidazol-4-yl)but-2-enoic	Asp 482, Asp 401, Lys 425	-51.556
ZINC01235906	6,6,7,7-tetrafluoro-3-oxabicyclo[3.2.0] heptane-2,4-di carboxylicacid	Asp 482, 2/ Lys 425	-6.482

TABLE 4.14: ZincID, Minimized energy, Scientific names of Compounds , Number of interactions and Interactive Residues for 2-isopropylmalate synthase

Zinc ID	Scientific names of Compounds	Number of interaction & Interacting Residues	Minimized Energy
ZINC40448986	H-Pro(4-N3). HCl (2S,4S)	Asp 401, 2/ Lys Asp 402, Asp 482	-8.691
ZINC49625635	2,3-dihydroimidazo [1,2-c]isoxazolo[4, 5-e]pyrimidine-9 -carboxylic acid	2/ Lys 425, Asp 401, Asp 482	-40.749
ZINC39134339	2,3-dihydro-6-quinoxa linecarboxylicacid	Asp 401, Asp 482, Lys 425	-12.214
ZINC38342322	2-[(2-aminoethyl)amino] pyridine-4-carboxylic acid	2/ Lys 425, Asp 401, Asp 402	-16.075

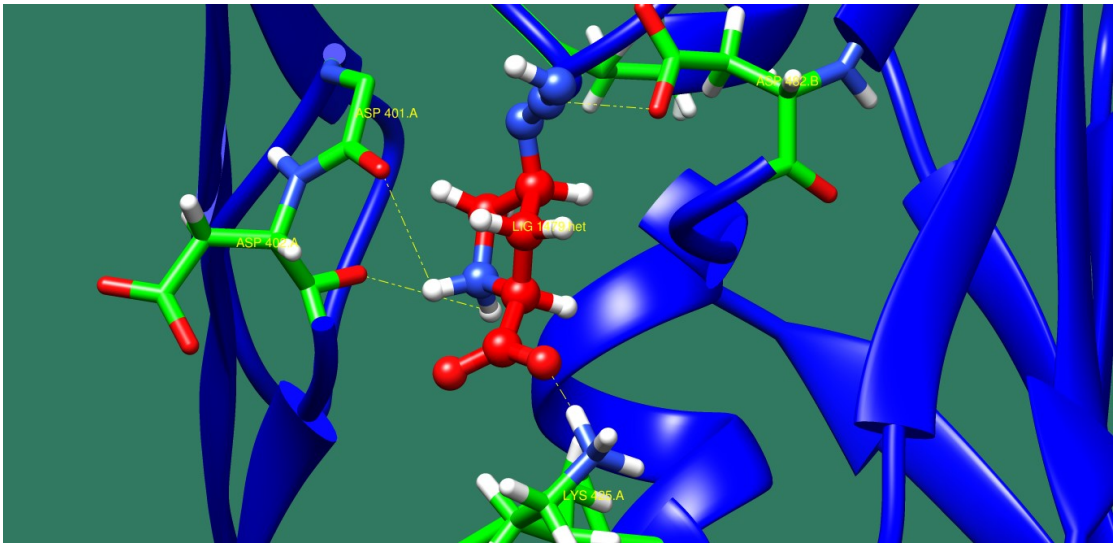

FIGURE 4.10: Interaction of 2-isopropylmalate synthase with ZINC40448986

Ribosome-binding factor A (rbfA) is cold shock adaptation protein which helps bacteria to grow at low temperature (10-20 ℃). This protein associates with 30S ribosomal subunit but do not associates with 70S ribosomes or polysomes. It

also interacts with 5'-terminal helix of 16S rRNA. During the cold shock adaptation several cold shock proteins are synthesized which allow the efficient translation processing of the mRNAs which facilitates the ribosome assembly that is required for the growth of bacteria(Huang et al., 2003). As this protein is found to be virulent and quite essential for bacteria so that it could act as potential drug target. The best interacting lead molecules are shown in table 4.15 along with ZincID, minimized energy, number of interactions and interacting residues. ZINC01235906 was predicted as top ranked molecule interacting with binding site residues lys24 and Arg77 (Figure 4.11).

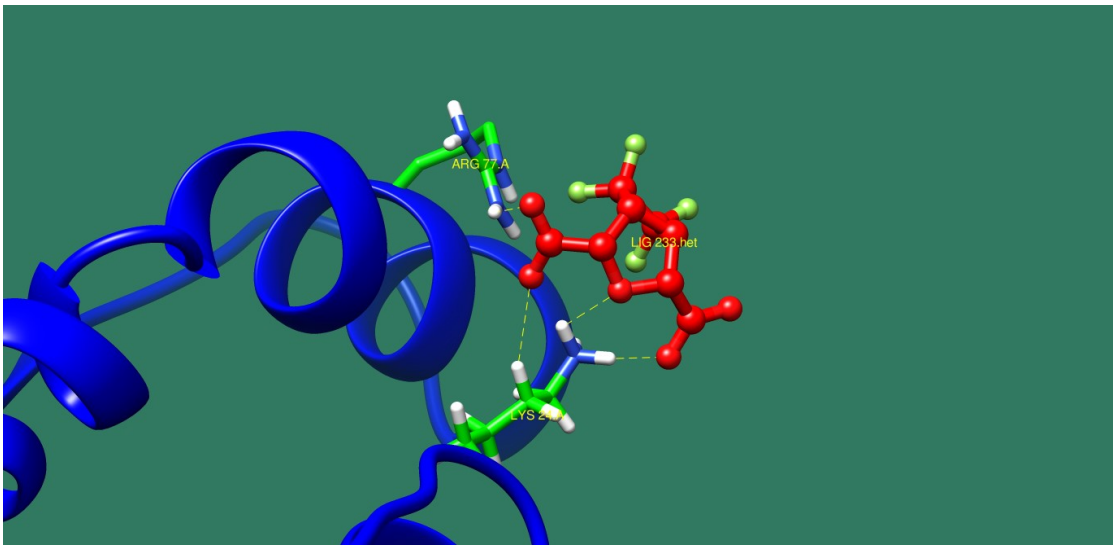

FIGURE 4.11: Interaction of Ribosome-binding factor A with ZINC01235906

TABLE 4.15: ZincID, Minimized energy, Scientific names of Compounds , Number of interactions and Interactive Residues for Ribosome-binding factor A

Zinc ID	Scientific names of Compounds	Number of interactions and Interacting Residues	Minimized Energy
ZINC149388367	Not known	Lys 70	-14.308
ZINC83235996	[(E)-(5-hydroxy-3-methyl-oxadiazol-3-ium-4-yl) methyleneamino]thiourea	Lys 24, Arg 77	-6.721

TABLE 4.15: ZincID, Minimized energy, Scientific names of Compounds , Number of interactions and Interactive Residues for Ribosome-binding factor A

Zinc ID	Scientific names of Compounds	Number of interactions and Interacting Residues	Minimized Energy
ZINC01235906	6,6,7,7-tetrafluoro-3-oxabi cyclo[3.2.0]heptane-2,4 -dicarboxylic acid	3/ lys 24, Arg 77	-17.173
ZINC00171258	3-(2,5-difluorophenyl)-2-hydroxy-6,7,8,9-tetrahydro -4H-pyrido[1,2-a] pyrimidin-4-one	Lys 24, Arg 77	-10.793
zinc00255388	2,3-dihydroimidazo[2,1-b] quinazolin-1-ium-5-olate	2/ Arg 26, Lys 24	-6.936
ZINC01532584	Prephenate	3/ Lys 24, Arg 77	-19.888
ZINC03872713	Glyphosate	Arg 77, Arg 81, 2/ Lys 24	-15.475
ZINC05185127	N-[(1-benzyl-4-piperidylidene) amino]-1H-tetrazol -5-amine	Asp 27, Lys 63	-22.165
ZINC03852636	BLAHol	Lys 24, Arg 77, Arg 26	-10.595
ZINC58386852	(2R)-1-methylsulfonyl-N-(4H-1,2,4-triazol-3-yl) piperidine-2-carboxamide	Thr 74, Lys 24, Arg77	-10.798

DNA-binding response regulator (DW662-02135) is a protein which mediates the change in cell according to the response in the environment. This protein is a part of two components regulatory system(TCS). Bacteria tends to change

its environment according to different levels regulation and expression of genes, expression of multiple operons and stress response and sporulation and cellular motility, cell aggregation and biofilm formation. All these levels are controlled by TCS from primarily to transcription, translations and post translation of regulation of genes and also through different type of protein-protein interaction and also its virulence. TCS consist of histidine kinases which sense the environmental signal and generates the response regulator.

TABLE 4.16: ZincID, Minimized energy, Scientific names of Compounds , Number of ineractions and Interactive Residues for DNA-binding response regulator

Zinc ID	Scientific names of Compounds	Number of interactions and Interacting Residues	Minimized Energy
ZINC22108884	3-[hydroxy-[2-(6-methyl -3-pyridyl)ethyl] phosphoryl]propanoic	2/ Lys 153, 2/ Lys 156, 2/ Arg 117	-27.838
ZINC27572262	2-(4-nitroimino-1,3,5 -triazinan-1-yl)ethyl	Arg 117, Lys 156, Lys 153, Lys 156, Gln 140	-19.566
ZINC31156942	(2S)-2-[(3,6-dihydroxy-5-methoxy-7-methyl-4-oxo-chromen-2-yl) amino]propanoic	Gly 138, Gln 140, Arg 117, 2/ Lys 156, Lys 153	-26.579
ZINC17353456	2-(3,6-Dioxo-1,2,3,6-tetrahydropyridazin-4-yl)acetic acid	Gln 140, 3/ Lys 156, 2/ Arg 117, Lys 153	-23.609
ZINC71777127	Not Known	2/ Lys 153, 2/Lys 156, Arg 117, His 74, Gln 140	-30.601

TABLE 4.16: ZincID, Minimized energy, Scientific names of Compounds , Number of ineractions and Interactive Residues for DNA-binding response regulator

Zinc ID	Scientific names of Compounds	Number of interactions and Interacting Residues	Minimized Energy
ZINC72271115	(2S)-2-cyano-3-imidazo [1,2-a]pyridin-2-yl-3-oxo-N-(2-pyridyl) propanamide	Ser 114, 2/ Lys 156, 2/ Lys 153	-25.49
ZINC01532584	Prephenate	Lys 153, 2/ Lys 156, Arg 117, His 74	-27.704
ZINC01618279	(E,1S)-1-hydroxy-3-phenyl-prop-2-ene-1 -sulfonate	Gly 152, Lys 153, 2/ Lys 156, Arg 117	-15.932
ZINC01673626	N-butyl phosphate	3/ Lys 156, 2/Lys 153	-29.47
ZINC38140720	medronic acid	His 74, Ser 114, Arg 117,3/ Lys 156, Lys 153,	-34.255

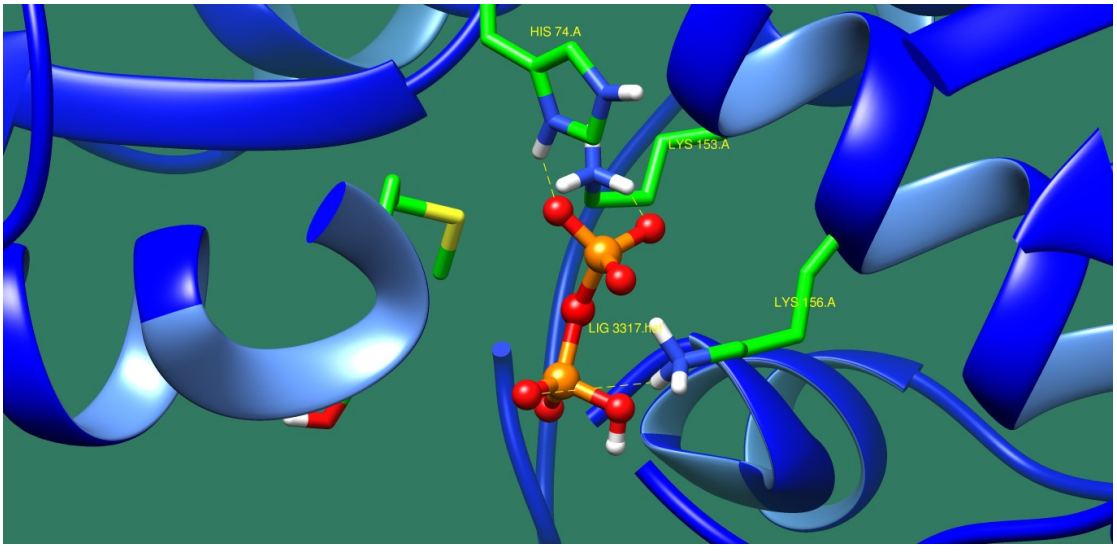

FIGURE 4.12: Interaction of DNA-binding response regulator with ZINC 38140720

This process is phosphorylated by the cognate histidine kinase and it also sometimes function as transcription regulator to regulate the expression of genes (Wang et al., 2007; Galperin, 2010). As this protein is non-homolog to human and also found to be essential and virulent so this protein could be a potential drug target against Sg. Table 4.16 presents best interacting lead molecules along with their ZincID, Interacting Residues, number of interactions and minimized energy. Binding site residues His74, Ser114, Arg117, Lys156 and Lys153 were predicted to interact with ZINC38140720 as shown in figure 4.12.

For each target protein, we were able to shortlist 10 lead molecules out of which 1 molecule was ranked on top. It would be appropriate to translate these in-silico findings into in-vitro and finally in-vivo to channelize the computational findings toward experimental validation.

Chapter 5

Conclusions and Recommendations

Streptococcus gallolyticus (Sg) is an opportunistic bacterium which causes infective endocarditis that is an inflammation of heart. This bacterium has developed antibiotic resistance towards the available drugs. So, therefore there is a need for novel therapeutics targets which could prevent the onset of this disease. For this we have used in-silico approach, in which first we identified the core genes through pan genome analysis. From these core genes we have identified the essential non-host homologous genes from subtractive genomic approach. Then we prioritize our drug targets and identified the potent lead compounds using protein-ligand interaction.

The first objective was the identification core genome of all strains of Sg. To achieve this objective total 7 strain of Sg sequenced were retrieved. Then the complete set of these strains was used to identify pan-genome. The complete set of strains represents pan-genome. This complete set consists of 3,242 genes. From which only those genes were selected which were common in all the strain which is called core genome. The total core genomes that have been identified were 1,138.

The second objective of this study was to perform subtractive genomic analysis. This objective was achieved by identifying the non-homologous and essential

genes. A total of 1,115 non-host homologous genes were identified. These non-host homologous were absent in the host (human). From these 1,115 non-host homologous genes 18 essential genes were identified which were important for Sg survival and its growth. These 18 essentials non-host homologous genes indicate that these genes are absent in host and are essential for the Sg can play an important role in causing the disease.

The third objective of this study was to prioritize our protein targets, identification of potent lead compounds using protein-ligand interaction. This objective was achieved through the prioritization of drug targets from different factors. The first factor was molecular weight, in which all proteins have low molecular weight consisting molecular weight less than 100 kDa. This helps the compounds to absorb easily in the membrane. Second factor was pathway analysis which shows that these targets play a vital role in survival of bacteria as some of these targets were involved in biosynthesis of peptidoglycan and some of them are heat shock proteins. Third factor was identification of virulent genes. When these targets were blast against virulent factor database, all proteins were found to be virulent. Then the molecular and biological processes were identified which provided information about their functional or biological role in Sg. Then subcellular localization was done which indicates that 12 were cytoplasmic proteins, 4 were membrane protein and 2 was extracellular protein. The 12 cytoplasmic could act as drug targets while the other proteins could act as vaccine targets. The docking of these drug targets was done which showed favorable interactions against compounds library. The selected compounds were retrieved from literature. Some of these targets are experimentally validated and already reported as drug/vaccine targets in another organism. All these interactions were bound at very low binding energy which indicates theses are stable molecules. These results indicate that these 18 drug and vaccine targets can be used for designing of new drugs/vaccine with low probability of side effects and can prevent the onset of this disease.

For the future work, the experimental validation of these targets is suggested to validate their role in survival and virulence of Sg. The laboratory experiments can ultimately result in commercial products in future.

Bibliography

[1] E. Pasquereau-Kotula, M. Martins, L. Aymeric, and S. Dramsi, "Significance of Streptococcus gallolyticus subsp. gallolyticus association with colorectal cancer," Frontiers in Microbiology, vol 9, no. 614 APR. Frontiers Media S.A., 03-Apr-2018.

[2] A. Harrington and Y. Tal-Gana, "Identification of Streptococcus gallolyticus subsp. gallolyticus (biotype I) competencestimulating peptide pheromone," J. Bacteriol., vol. 200, no. 14, Jul. 2018.

[3] N. Takamura, T. Kenzaka, K. Minami, and M. Matsumura, "Infective endocarditis caused by Streptococcus gallolyticus subspecies pasteurianus and colon cancer," BMJ Case Rep.,vol. 2014, no. bcr2013203476 May 2014.

[4] M. E. Hensler, "Streptococcus gallolyticus, infective endocarditis, and colon carcinoma: New light on an intriguing coincidence," Journal of Infectious Diseases, vol. 203, no. 8, pp. 1040–1042, 15-Apr-2011.

[5] C. Rusniok et al., "Genome Sequence of Streptococcus gallolyticus: Insights into Its Adaptation to the Bovine Rumen and Its Ability To Cause Endocarditis †," J. Bacteriol., vol. 192, no. 8, pp. 2266–2276, 2010.

[6] D. Hinse et al., "Complete genome and comparative analysis of Streptococcus gallolyticus subsp. gallolyticus, an emerging pathogen of infective endocarditis," BMC Genomics, vol. 12,no. 400 Aug. 2011.

[7] B. C. M. and J. E. Moore, "Emerging Issues in Infective Endocarditis - Volume 10, Number 6—June 2004 - Emerging Infectious Diseases journal - CDC," Emerg. Infect. Dis. J., vol. Volume 10, no. Number 6, pp. 1110–1116, 2004.

[8] M. F. Tripodi, R. Fortunato, R. Utili, M. Triassi, and R. Zarrilli, "Molecular epidemiology of Streptococcus bovis causing endocarditis and bacteraemia in Italian patients," Clin. Microbiol. Infect., vol. 11, no. 10, pp. 814–819, 2005.

[9] B. Hoen et al., "Emergence of endocarditis due to group D streptococci: Findings derived from the merged database of the International Collaboration on Endocarditis," Eur. J. Clin. Microbiol. Infect. Dis., vol. 24, no. 1, pp. 12–16, Jan. 2005.

[10] "Contemporary Challenges in Endocarditis - Google Books." [Online]. Available: https://books.google.com.pk/books [Accessed: 31-Oct-2019].

[11] V. Vilcant and O. Hai, Endocarditis, Bacterial. 2018.

[12] M.-C. Tomescu, "Streptococcus gallolyticus spontaneous infective endocarditis on native valves, in a diabetic patient Correspondence to." Medicina in Evolutie, vol. 3, pp. 323-328, Jan 2014.

[13] J. R. McDonald, "Acute Infective Endocarditis," Infect. Dis. Clin. North Am., vol. 23, no. 3, pp. 643–664, Sep. 2009.

[14] E. A. Ashley and J. Niebauer, Cardiology explained Remedica Explained Series. 2004.

[15] J. A. Satué Bartolomé and M. Alonso Sanz, "Streptococcus gallolyticus: a new name for a well-known old organism." Archivos de Medicina, vol. 5, no. 1, Jan. 2009

[16] H. Grubitzsch, T. Christ, C. Melzer, M. Kastrup, S. Treskatsch, and W. Konertz, "Surgery for prosthetic valve endocarditis: Associations between morbidity, mortality and costs," Interact. Cardiovasc. Thorac. Surg., vol. 22, no. 6, pp. 784–791, Jun. 2016.

[17] S. B. Jamal et al., "An integrative in-silico approach for therapeutic target identification in the human pathogen Corynebacterium diphtheriae," PLoS One, vol. 12, no. 10, Oct. 2017.

[18] J. Sillanpää et al., "A collagen-binding adhesin, Acb, and ten other putative MSCRAMM and pilus family proteins of Streptococcus gallolyticus subsp. gallolyticus (Streptococcus bovis group, biotype I)," J. Bacteriol., vol. 191, no. 21, pp. 6643–6653, Nov. 2009.

[19] J. Sillanpää, S. R. Nallapareddy, K. V. Singh, M. J. Ferraro, and B. E. Murray, "Adherence characteristics of endocarditis-derived Streptococcus gallolyticus ssp. gallolyticus (Streptococcus bovis biotype I) isolates to host extracellular matrix proteins," FEMS Microbiol. Lett., vol. 289, no. 1, pp. 104–109, Dec. 2008.

[20] "Streptococcus gallolyticus." [Online]. Available: http://microbe-canvas.com/Bacteria.php?p=1175. [Accessed: 07-Feb-2020].

[21] L. M. Baddour et al., "Infective endocarditis in adults: Diagnosis, antimicrobial therapy, and management of complications: A scientific statement for healthcare professionals from the American Heart Association," Circulation, vol. 132, no. 15. Lippincott Williams and Wilkins, pp. 1435–1486, 13-Oct-2015.

[22] J. Isenring et al., "Streptococcus gallolyticus subsp. Gallolyticus endocarditis isolate interferes with coagulation and activates the contact system," Virulence, vol. 9, no. 1, pp. 248–261, Dec. 2018.

[23] T. L. Holland, L. M. Baddour, A. S. Bayer, B. Hoen, J. M. Miro, and V. G. Fowler, "Infective endocarditis," Nat. Rev. Dis. Prim., vol. 2,no. 23, pp. e394–e434, Sep. 2016.

[24] "Clinical Profile and Outcome of Infective Endocarditis at the Aga Khan University Hospital — Insight Medical Publishing." [Online]. Available: http://internalmedicine.imedpub.com [Accessed: 07-Dec-2019].

[25] M. Tariq, M. Alam, G. Munir, M. A. Khan, and R. A. Smego, "Infective endocarditis: A five-year experience at a tertiary care hospital in Pakistan," Int. J. Infect. Dis., vol. 8, no. 3, pp. 163–170, 2004.

[26] U. Shahid, H. Sharif, J. Farooqi, B. Jamil, and E. Khan, "Microbiological and clinical profile of infective endocarditis patients: an observational study experience from tertiary care center Karachi Pakistan," J. Cardiothorac. Surg., vol. 13, no. 1, p. 94, Dec. 2018.

[27] M. A. Iqbal, M. Fahim, and N. Khan, "Im-hospital outcome of patients with infective endocarditis", Khyber Medical University Journal, vol. 7, pp. 4–7, Jul. 2015.

[28] E. Durante-Mangoni et al., "Current features of infective endocarditis in elderly patients: Results of the international collaboration on endocarditis prospective cohort study," Arch. Intern. Med., vol. 168, no. 19, pp. 2095–2103, Oct. 2008.

[29] "JPMA - Journal Of Pakistan Medical Association." [Online]. Available: https://jpma.org.pk/article-details/1895. [Accessed: 07-Dec-2019].

[30] L. Carlos Guimaraes et al., "Inside the Pan-genome - Methods and Software Overview," Curr. Genomics, vol. 16, no. 4, pp. 245–252, May 2015.

[31] H. Tettelin, ... V. M.-P. of the, and undefined 2005, "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial 'pan-genome,'" Natl. Acad Sci,vol. 210, no. 6, pp. 1265-1281, Jun. 2013.

[32] "What is a Pan Genome? - www.10wheatgenomes.com." [Online]. Available: http://www.10wheatgenomes.com/what-is-a-pan-genome/. [Accessed: 07-Feb-2020].

[33] J. Blom et al., "EDGAR 2.0: an enhanced software platform for comparative gene content analyses," Nucleic Acids Res., vol. 44, no. W1, pp. W22–W28, Jul. 2016.

[34] L. Li, C. J. Stoeckert, and D. S. Roos, "OrthoMCL: Identification of ortholog groups for eukaryotic genomes," Genome Res., vol. 13, no. 9, pp. 2178–2189, Sep. 2003.

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

[36] A. Ali et al., "Pan-genome analysis of human gastric pathogen H. pylori: Comparative genomics and pathogenomics approaches to identify regions associated with pathogenicity and prediction of potential core therapeutic targets," Biomed Res. Int., vol. 2015,no. 143,Oct. 2015.

[37] S. S. Hassan et al., "Proteome scale comparative modeling for conserved drug and vaccine targets identification in Corynebacterium pseudotuberculosis," BMC Genomics, vol. 15, no. S7, p. S3, Oct. 2014.

[38] A. K. Jaiswal, S. Tiwari, S. B. Jamal, D. Barh, V. Azevedo, and S. C. Soares, "An in silico identification of common putative vaccine candidates against treponema pallidum: A reverse vaccinology and subtractive genomics based approach," Int. J. Mol. Sci., vol. 18, no. 2, Feb. 2017.

[39] B. Rathi, A. N. Sarangi, and N. Trivedi, "Genome subtraction for novel target definition in Salmonella typhi.," Bioinformation, vol. 4, no. 4, pp. 143–50, Oct. 2009.

[40] Z. Zhang and Q. Ren, "Why are essential genes essential? - The essentiality of Saccharomyces genes," Microbial Cell, vol. 2, no. 8. Shared Science Publishers OG, pp. 280–287, 01-Aug-2015.

[41] C. Peng, Y. Lin, H. Luo, and F. Gao, "A comprehensive overview of online resources to identify and predict bacterial essential genes," Front. Microbiol., vol. 8, no. 2331, Nov. 2017.

[42] Y.-N. Ye, Z.-G. Hua, J. Huang, N. Rao, and F.-B. Guo, "CEG: a database of essential gene clusters," BMC Genomics, vol. 14, no. 1, p. 769, Nov. 2013.

[43] R. Zhang, "DEG: a database of essential genes," Nucleic Acids Res., vol. 32, no. 90001, pp. 271D – 272, Jan. 2004.

[44] K.-H. Thierauch, "Small Molecule Drugs," in Encyclopedia of Cancer, Berlin, Heidelberg: Springer Berlin Heidelberg,vol.1 , pp. 3448–3451 , 2011.

[45] "ExPASy - ProtParam documentation." [Online]. Available: https://web.expasy.org/protparam/protparam-doc.html. [Accessed: 08-Feb-2020].

[46] "Gene Ontology overview." [Online]. Available: http://geneontology.org/docs/ontology-documentation/. [Accessed: 06-Feb-2020].

[47] R. Apweiler, "The Universal Protein resource (UniProt)," Nucleic Acids Res., vol. 36, no. SUPPL. 1, p. D190, Jan. 2008.

[48] M. S. Scott, S. J. Calafell, D. Y. Thomas, and M. T. Hallett, "Refining Protein Subcellular Localization," PLoS Comput. Biol., vol. 1, no. 6, p. e66, 2005.

[49] C.-S. Yu, C.-J. Lin, and J.-K. Hwang, " Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n -peptide compositions ," Protein Sci., vol. 13, no. 5, pp. 1402–1406, May 2004.

[50] J. L. Gardy et al., "PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria," Nucleic Acids Res., vol. 31, no. 13, pp. 3613–3617, Jul. 2003.

[51] O. Emanuelsson, H. Nielsen, S. Brunak, and G. Von Heijne, "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence," J. Mol. Biol., vol. 300, no. 4, pp. 1005–1016, Jul. 2000.

[52] S. Hua and Z. Sun, "Support vector machine approach for protein subcellular localization prediction,"Bioinformatics, vol. 17, no. 8, pp. 721-728, Aug. 2001.

[53] H. Ma and H. Zhao, "Drug target inference through pathway analysis of genomics data," Advanced Drug Delivery Reviews, vol. 65, no. 7. NIH Public Access, pp. 966–972, 30-Jun-2013.

[54] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: New perspectives on genomes, pathways, diseases and drugs," Nucleic Acids Res., vol. 45, no. D1, pp. D353–D361, Jan. 2017.

[55] L.-L. Zheng et al., "A Comparison of Computational Methods for Identifying Virulence Factors," PLoS One, vol. 7, no. 8, p. e42517, Aug. 2012.

[56] D. Che, M. Hasan, and B. Chen, "Identifying Pathogenicity Islands in Bacterial Pathogenomics Using Computational Approaches," Pathogens, vol. 3, no. 1, pp. 36–56, Jan. 2014.

[57] L. Chen et al., "VFDB: A reference database for bacterial virulence factors," Nucleic Acids Res., vol. 33, no. DATABASE ISS., Jan. 2005.

[58] M. Cammisa, A. Correra, G. Andreotti, and M. V. Cubellis, "Identification and analysis of conserved pockets on protein surfaces," BMC Bioinformatics, vol. 14, no. SUPPL7, p. S9, Apr. 2013.

[59] "Chemical databases/resources." [Online]. Available: http://crdd.osdd.net/chemdatabase.php. [Accessed: 10-Feb-2020].

[60] M. S. Saddala and P. J. Adi, "Discovery of small molecules through pharmacophore modeling, docking and molecular dynamics simulation against Plasmodium vivax Vivapain-3 (VP-3)," Heliyon, vol. 4, no. 5, May 2018.

[61] A. M. Dar and S. Mir, "Molecular Docking: Approaches, Types, Applications and Basic Challenges," J. Anal. Bioanal. Tech., vol. 08, no. 02, pp. 1–3, Mar. 2017.

[62] J. de Ruyck, G. Brysbaert, R. Blossey, and M. F. Lensink, "Molecular docking as a popular tool in drug design, an in silico travel," Advances and Applications in Bioinformatics and Chemistry, vol. 9, no. 1. Dove Medical Press Ltd, 2016.

[63] G. B. Ray and J. W. Cook, "Molecular modeling of heme proteins using MOE: Bio-inorganic and structure-function activity for undergraduates," Biochem. Mol. Biol. Educ., vol. 33, no. 3, pp. 194–201, May 2005.

[64] E. Trost et al., "Pangenomic study of Corynebacterium diphtheriae that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia," J. Bacteriol., vol. 194, no. 12, pp. 3199–3215, Jun. 2012.

[65] H. Luo, Y. Lin, F. Gao, C. T. Zhang, and R. Zhang, "DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements," Nucleic Acids Res., vol. 42, no. D1, Jan. 2014.

[66] F. Agüero et al., "Genomic-scale prioritization of drug targets: The TDR Targets database," Nat. Rev. Drug Discov., vol. 7, no. 11, pp. 900–907, 2008.

[67] S. I. Mondal et al., "Identification of potential drug targets by subtractive genome analysis of Escherichia coli O157:H7: An in silico approach," Adv. Appl. Bioinforma. Chem., vol. 8, no. 1, pp. 49–63, 2015.

[68] A. Wadood, S. B. Jamal, M. Riaz, and A. Mir, "Computational analysis of benzofuran-2-carboxlic acids as potent Pim-1 kinase inhibitors," Pharm. Biol., vol. 52, no. 9, pp. 1170–1178, 2014.

[69] H. S. Sader, D. J. Farrell, and R. N. Jones, "Antimicrobial activity of daptomycin tested against gram-positive strains collected in European hospitals: Results from 7 years of resistance surveillance (2003-2009)," J. Chemother., vol. 23, no. 4, pp. 200–206, 2011.

[70] M. Nielsen, C. Lundegaard, O. Lund, and T. N. Petersen, "CPHmodels-3.0—remote homology modeling using structure-guided sequence profiles" Nucleic Acids Res., vol. 38, no. suppl2, pp. W576–W581, Jul. 2010.

[71] P. Fariselli, I. Rossi, E. Capriotti, and R. Casadio, "The WWWH of remote homolog detection: The state of the art," Brief. Bioinform., vol. 8, no. 2, pp. 78–87, Dec. 2006.

[72] P. Batut and T. R. Gingeras, "RAMPAGE: Promoter activity profiling by paired-end sequencing of 5'-complete cDNAs," Curr. Protoc. Mol. Biol., vol. 104, no. 1, pp. 25B-11, oct. 2013.

[73] J. Hu, L. Zhao, and M. Yang, "A GntR family transcription factor positively regulates mycobacterial isoniazid resistance by controlling the expression of a putative permease," BMC Microbiol., vol. 15, no. 1, p. 214, Dec. 2015.

[74] B. Hillerich and J. Westpheling, "A New GntR Family Transcriptional Regulator in Streptomyces coelicolor Is Required for Morphogenesis and Antibiotic Production and Controls Transcription of an ABC Transporter in Response to Carbon Source," J. Bacteriol., vol. 188, no. 21, pp. 7477–7487, 2006.

[75] V. Vindal, K. Suma, and A. Ranjan, "GntR family of regulators in Mycobacterium smegmatis: A sequence and structure based characterization," BMC Genomics, vol. 8, Aug. 2007.

[76] R. Adhikari, D. Singh, M. Chandravanshi, A. Dutta, and S. P. Kanaujia, "UgpB, a periplasmic component of the UgpABCE ATP-binding cassette transporter, predominantly follows the Sec translocation pathway," Meta Gene, vol. 13, pp. 129–139, Sep. 2017.

[77] R. A. Wing, S. Bailey, and T. A. Steitz, "Insights into the Replisome from the Structure of a Ternary Complex of the DNA Polymerase III α-Subunit," J. Mol. Biol., vol. 382, no. 4, pp. 859–869, Oct. 2008.

[78] J. Liu et al., "Crystal structure of a heat-inducible transcriptional repressor HrcA from Thermotoga maritima: Structural insight into DNA binding and dimerization," J. Mol. Biol., vol. 350, no. 5, pp. 987–996, Jul. 2005.

[79] A. Mitra, Y. H. Ko, G. Cingolani, and M. Niederweis, "Heme and hemoglobin utilization by Mycobacterium tuberculosis," Nat. Commun., vol. 10, no. 1, Dec. 2019.

[80] S. Létoffé, P. Delepelaire, and C. Wandersman, "The housekeeping dipeptide permease is the Escherichia coli heme transporter and functions with two optional peptide binding proteins," Proc. Natl. Acad. Sci. U. S. A., vol. 103, no. 34, pp. 12891–12896, Aug. 2006.

[81] J. Zhang et al., "Snapshots of catalysis: Structure of covalently bound substrate trapped in Mycobacterium tuberculosis thiazole synthase (ThiG)," Biochem. Biophys. Res. Commun., vol. 497, no. 1, pp. 214–219, Feb. 2018.

[82] C. R. Machado, R. L. de Oliveira, S. Boiteux, U. M. Praekelt, P. A. Meacock, and C. F. Menck, "Thi1, a thiamine biosynthetic gene in Arabidopsis thaliana, complements bacterial defects in DNA repair.," Plant Mol. Biol., vol. 31, no. 3, pp. 585–93, Jun. 1996.

[83] J. B. Thoden, X. Huang, F. M. Raushel, and H. M. Holden, "Carbamoyl-phosphate synthetase: Creation of an escape route for ammonia," J. Biol. Chem., vol. 277, no. 42, pp. 39722–39727, Oct. 2002.

[84] D. Shi, L. Caldovic, and M. Tuchman, "Sources and fates of carbamyl phosphate: A labile energy-rich molecule with multiple facets," Biology, vol. 7, no. 2. ,pp. 34, MDPI AG, 01-Jun-2018.

[85] G. A. Prosser and L. P. S. De Carvalho, "Kinetic mechanism and inhibition of Mycobacterium tuberculosis d-alanine: D-alanine ligase by the antibiotic d-cycloserine," FEBS J., vol. 280, no. 4, pp. 1150–1166, Feb. 2013.

[86] J. H. Lee et al., "Crystal structure of the apo form of D-alanine: D-alanine ligase (Ddl) from Thermus caldophilus: A basis for the substrate-induced conformational changes," Proteins Struct. Funct. Bioinforma., vol. 64, no. 4, pp. 1078–1082, Jun. 2006.

[87] A. P. Kuzin, T. Sun, J. Jorczak-Baillass, V. L. Healy, C. T. Walsh, and J. R. Knox, "Enzymes of vancomycin resistance: The structure of D-alanine-D-lactate ligase of naturally resistant Leuconostoc mesenteroides," Structure, vol. 8, no. 5, pp. 463–470, May 2000.

[88] N. Friedland et al., "Domain orientation in the inactive response regulator Mycobacterium tuberculosis MtrA provides a barrier to activation," Biochemistry, vol. 46, no. 23, pp. 6733–6743, Jun. 2007.

[89] H. Szurmant and J. A. Hoch, "Statistical analyses of protein sequence alignments identify structures and mechanisms in signal activation of sensor histidine kinases," Mol. Microbiol., vol. 87, no. 4, pp. 707–712, Feb. 2013.

[90] R. Agrawal and D. K. Saini, "Rv1027c-Rv1028c encode functional KdpDE two - Component system in Mycobacterium tuberculosis," Biochem. Biophys. Res. Commun., vol. 446, no. 4, pp. 1172–1178, Apr. 2014.

[91] Z. N. Freeman, S. Dorus, and N. R. Waterfield, "The KdpD/KdpE Two-Component System: Integrating K+ Homeostasis and Virulence," PLoS Pathog., vol. 9, no. 3, 2013.

[92] G. M. Amera, R. J. Khan, A. Pathak, A. Kumar, and A. K. Singh, "Structure based in-silico study on UDP-N-acetylmuramoyl-L-alanyl-D-glutamate-2,6-diaminopimelate ligase (MurE) from Acinetobacter baumannii as a drug target against nosocomial infections," Informatics Med. Unlocked, vol. 16, no. 10, pp. 100216, Jan. 2019.

[93] R. Fernandez-López, R. Ruiz, F. de la Cruz, and G. Moncalián, "Transcription factor-based biosensors enlightened by the analyte," Frontiers in Microbiology, vol. 6, no. 648 JUL. Frontiers Research Foundation, 2015.

[94] Q. Li, C. Li, L. Xie, C. Zhang, Y. Feng, and J. Xie, "Characterization of a putative ArsR transcriptional regulator encoded by Rv2642 from Mycobacterium tuberculosis," Journal of Biomolecular Structure and Dynamics, vol. 35, no. 9. Taylor and Francis Ltd., pp. 2031–2039, 04-Jul-2017.

[95] G. Spraggon et al., "Crystal structure of an Udp-n-acetylmuramate-alanine ligase MurC (TM0231) from Thermotoga maritima at 2.3 Å resolution," Proteins Struct. Funct. Genet., vol. 55, no. 4, pp. 1078–1081, Apr. 2004.

[96]. Guo, D., et al., Reactive oxygen species-induced cytotoxic effects of zinc oxide nanoparticles in rat retinal ganglion cells. Toxicology in Vitro, vol. 27 no. 2, p. 731-738,Jul 2013.

[97]. E. F. Pettersen et al., "UCSF Chimera - A visualization system for exploratory research and analysis," J. Comput. Chem., vol. 25, no. 13, pp. 1605–1612, Oct. 2004, doi: 10.1002/jcc.20084.