Characterization of Druggable Genome for the Identification of Potential Therapeutic Candidates in Human Pathogen: *Vibrio Cholerae*

By

Iqra Ahmad

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-i-Azam University Islamabad, Pakistan

2016

Characterization of Druggable Genome for the Identification of Potential Therapeutic Candidates in Human Pathogen: *Vibrio Cholerae*

A Thesis submitted in the partial fulfillment of the requirements for the degree of

Master of Philosophy in Bioinformatics

By

Iqra Ahmad

Supervisor

Dr. Syed Sikander Azam

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-i-Azam University Islamabad, Pakistan

2016

Dedication

To
My Affectionate Parents
"Mr. & Mrs. Iftikhar Ahmad"
Who Always Raise Their Hands to Pray for Me!

To
My Compassionate Parents-in-Law
"Mr. & Mrs. Izhar Hussain"
For Their Love, Support and Encouragement!

And
To My Soulmate
'Khurram Izhar"
Who is the World for Me
Without Him the Present Destination
Would Have Been a Mere Dream!

DECLARATION

I hereby solemnly declare that the work **"Characterization of Druggable Genome for the Identification of Potential Therapeutic Candidates in Human Pathogen: *Vibrio Cholerae"*** presented in the following thesis is my own effort, except where otherwise acknowledged and that the thesis is my own composition. No part of the thesis has been previously presented for any other degree.

Dated: _____ Iqra Ahmad

Acknowledgements

In the name of ALLAH (SWT), The Most Bounteous and The Most Merciful, all praises to ALLAH (SWT) for strength and His blessing in the accomplishment of this thesis. No words can express enough gratitude for His knowledge and His wisdom.

I would like to express my gratitude to Dr. Syed Sikander Azam for being an outstanding advisor and excellent supervisor. His constant encouragement, support, and invaluable suggestions made this work successful. I am indebted and thankful to him.

I would like to acknowledge Chairperson Dr. Sajid Rashid for the academic support.

I am thankful to Prof. Dr. Klaus R. Liedl, for lending us computational help for the extension of simulations. My sincere thanks to our Ph.D student Saad Raza for the Perl scripts, his help was really valuable for this work. I am indebted to Computational biology group members Asma Abro, Samra Wajid Abbasi, Amen Shamim, Seemab Khurshid, Hira Jabeen, Sanam Javed, Nimara Sabir, Noor Ul Ain Sajid Mughal, Sundus Iqbal, Fouzia Shaheen, Gul Sanober and Farhan ul Haq for providing a stimulating and constructive environment in lab.

I would like to pay high regards to my parents, mother-in-law Habiba Yasmin, my husband Khurram Izhar, my siblings Azka, Osamah and Umamah, my sisters-in-law Ailya and Sara, and my friends Noor, Sundus, Abida and Shabana for their sincere encouragement and unconditional love throughout my research work. I owe everything to them.

Iqra Ahmad

Table of Contents

Table of Contents

List of Abbreviations

Three-Dimensional	3D
Assisted Model Building with Energy Refinement	AMBER
Base pair	Bp
Basic Local Alignment Search Tool	BLAST
Beta Factor	β-Factor
BRaunschweig ENzyme Database	BRENDA
Broyden-Fletcher-Goldfarb-Shanno	BFGS
Computer Aided Drug Design	CADD
Cluster Database at High Identity with Tolerance	CD-HIT
Subcellular Localization	CELLO
Cholera Toxin	CT
Database of Essential Genes	DEG
Enzyme Commission	EC
Expectation Value	E- Value
Discrete Optimized Protein Energy	DOPE
Genetic Algorithm	GA
Genetic Optimization for Ligand Docking	GOLD
KEGG Automatic Annotation Server	KASS
Kyoto Encyclopedia of Genes and Genomes	KEGG

List of Abbreviations

Molecular Dynamics	MD
Multi Drug Resistant	MDR
National Center for Biotechnology Information	NCBI
Nanosecond	ns
National Institute of Health	NIH
Open Reading Frames	ORFs
Position Specific Iterative-BLAST	PSI-BLAST
Process Trajectory	PTRAJ
Protein Data Bank	PDB
Radius of Gyration	Rg
Root Mean Square Deviation	RMSD
Root Mean Square Fluctuation	RMSF
Simulated Annealing with NMR Derived Energy Restraints	SANDER
Support Vector Machines	SVMs
Three-Point Transferable Intermolecular Potential	TIP3P
Toxin-Coregulated Pilus	TCP
Universal Protein Resource Knowledge Base	UniProtKB
Vibrio cholerae	*V. cholerae*
Visual Molecular Dynamics	VMD
World Health Organization	WHO

List of Figures

List of Figures

List of Tables

Abstract

Vibrio cholerae is the etiologic agent of the diarrheal disease cholera. *Vibrios* are Gram negative rod shape bacteria. This water-borne pathogen is attained through drinking of tainted water or eating contaminated food. The horizontal gene transfer of the virulence genes and the pathogenicity islands is the prominent cause of the emergence of new strains of *V. cholerae* and this instigates the research towards the identification of novel drugs. The objective of current study is to characterize and identify the common potential therapeutic drug targets against *V. cholerae* strains namely O395, LMA3984-4 and IEC224 by applying hierarchal *in silico* subtractive genomic approach accompanied by molecular docking and molecular dynamic (MD) simulation studies. After successful screening of druggable candidates, on the basis of set parameters, a potential drug target candidate vibE was selected. VibE is crucial for the synthesis of vibrobactin which belongs to biosynthesis of siderophore pathway. Siderophores have applications in medicine for antibiotics for improved drug targeting. Therefore, it is one of the most viable candidates for drug development. To this aim molecular modeling was carried out to gain insights into active site and modeling was performed via MODELLER and web servers. Best modeled structure was selected and used for molecular docking studies. Total 106 compounds library were prepared for docking and compound 103 was the best docked compound with a GOLDScore of 75.7. Moreover, time dependent dynamic behaviour of docked complexes were analyzed using MD simulation studies. MD trajectories analysis revealed the flexibility of loop region to stabilize the binding of ligand and target protein and hydrogen bonding pattern was also rearranged. These conformational changes suggested the potential of compound 103 to act as lead compound.

1 Introduction

1.1 The Genus *Vibrio*

Vibrios are Gram negative, rod shaped bacteria and fall under the order of *Gammaproteobacteria.* They are motile and use flagellum to locomote. They perform metabolism by fermentation or respiration process. They are largely halophilic with the exception of few species, which are nonhalophilic and their habitat is marine water, thus, they are frequently involved in seafood borne diseases (Gopal et al., 2005).

1.2 *Vibrio cholerae*

V. cholerae is the causative agent of the diarrheal disease cholera. *V. cholerae* is a waterborne pathogen that is attained through drinking of polluted water or eating contaminated food. *V. cholerae* species comprises of both pathogenic and nonpathogenic strains that only differ in their virulence gene content (Faruque, Albert and Mekalanos, 1998). The horizontal gene transfer of the virulence genes and pathogenicity islands is the prominent cause of the emergence of new strains of *V. cholerae* (Kovach, Shaffer and Peterson, 1996).

1.2.1 *V. cholerae* Associated Mortality

As cholera is a primordial disease, and there are numerous evidences tracing back to ancient times (Mekalanos, Rubin and Waldor, 1997). Until 1854 the mechanism of the spread of cholera amongst individuals was not learnt. The English physician John Snow stated that the spread of cholera was associated with consumption of drinking water. Cholera became the first reported disease during an epidemic in New York in 1866 (Duffy, 1971). Shortly after Snow's innovation of the source of cholera, Filippo Pacini, an Italian pathologist, proposed curved organisms in the luminal contents of intestines from patients who had cholera (Van Heyningen and Seal, 1983). According to WHO, it is estimated that there are 1.4 to 4.3 million cases, and 28 000 to 142 000 deaths worldwide due to cholera every year. The short incubation period of 2 hours to 5 days, is one of the factors that

triggers the potentially explosive pattern of epidemics (Daniels et al., 2000). These infections are not limited to America but other continents like Asia and Europe are equally affected (Shimada et al., 1994). The first six pandemics, all were identified in the Indian subcontinent and were caused by the classical biotype of the O1 of *V. cholerae* (Albert et al., 1993).

1.2.2 Multiple-Drug Resistance

Multiple drug resistance is a property present in pathogens (Piddock, 2006). *V. cholerae* shows resistance against antibiotics including chloramphenicol, tetracycline, sulfanilamide, furazoidone, trimethoprim-sulphamethoxazole, and erythromycin due to conjugation process with drug resistant plasmids from other *Vibrio* pathogenic strains (Hayashi et al., 1982). The genome sequence of *V. cholera* comprises a single copy of the cholera toxin (CT) genes located on chromosome 1. The ctxAB genes encrypt for the A and B subunits of CT and are titled as CTXΦ. Toxin-coregulated pilus (TCP) is the receptor for entry of CTX (Waldor and Mekalanos, 1996), and the TCP gene cluster is present on chromosome 1. Like the structural genes for CT and TCP, the regulatory gene, toxR, controls their expression (Lee et al., 1999). In contrast to CTX, RTX and well-known intestinal colonization factor, is located on chromosome 2 in *V. cholerae* (Lin et al., 1999).

1.2.3 Genomic Features

To date 490 different strains of *V. cholerae* have been found. A few, belonging to the serotype 01 are virulent human pathogens, reported till 2015 while only fifteen strains are completely sequenced. Sequenced data is available on National Center for Biotechnology Information (NCBI) *http://www.ncbi.nlm.nih.gov/genome/genomes/505*. The genome entails two circular chromosomes of 2,961,146 (chromosome 1) and 1,072,314 (chromosome 2) base pairs (bp). The average GC content of both chromosomes is 47.7% and 46.9%, respectively (Heidelberg et al., 2000). The total predicted open reading frames (ORFs) are 3,885. (Slamti et al., 2007). Genes that are required for growth, are located on chromosome 1. While genes that are present only on chromosome 2, are also supposed to be needed for normal cell functioning and play important roles in metabolic pathways (Yamaichi et al.,

1999). This study implies on exploring the genome of three strains of *V. cholerae* for identifying potential drug target. Three strains selected are O395, LMA3984-4 and IEC224. *Vibrio cholerae* O395 being one of the most commonly cholera causing strain, was first isolated from India and is reported worldwide. While LMA3984-4 strain was first isolated from Amazon environment and IEC224 was isolated from a cholera outbreak in the Brazil (de Sá Morais, 2012). Detailed genomic features of the selected strains are given in the Table 1.1.

Table 1.1. *Genomic features of O395, LMA3984-4 and IEC224.*

Strain Name	**Size (Mb)**	**GC%**	**Genes**	**Protein**
O395	4.13232	47.56	3917	3747
LMA3984-4	3.73872	47.97	3471	3076
IEC224	4.07959	47.49	3824	3677

1.2.3.1 Resistance Mechanism

Vibrio cholerae resistance mechanism is based on genes including SXT which is representative of a family of conjugative-transposon-like mobile genetic elements that encode multiple antibiotic resistance genes. In recent years, SXT-related conjugative, self-transmissible integrating elements have become widespread in Asian *V. cholerae*. Two SXT loci, designated *setC* and *setD*, were found to encode regulators that activate the transcription of genes required for SXT excision and transfer (Beaber, Hochhut and Waldor, 2002).

1.3 Applied *In silico* Approach

Numerous *in silico* approaches were adopted in the current study (Figure 1.1). These approaches were divided in following modules genome subtraction, drug target selection, comparative homology modeling, molecular docking and molecular dynamic simulations.

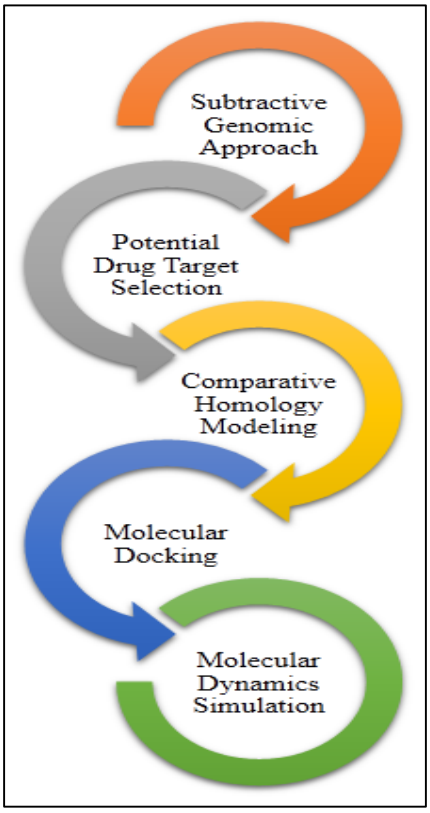

Figure 1. 1 *In silico steps adopted in the current study.*

1.3.1 Subtractive Genomic Approach

With the passage of time genomic and proteomic data is increasing since the accessibility of high throughput techniques is growing. Novel *in silico* approaches have been established to scrutinize the data and trade the orthodox procedures. Conventional drug discovery procedures were time consuming and tedious, whereas, *in silico* approach that is computer aided drug design (CADD) is far more superior and accurate process (Franklin, 2009). Subtractive genomic approach is a sequential procedure to identify pathogen vital proteins

and then use that protein as target for potential drug identification (Barh et al., 2011). This approach utilizes various databases, softwares and tools. Using this approach many pathogens have been used to locate potential drug targets for example *Pseudomonas aeruginosa, Helicobacter pylori* (Dutta et al., 2006), *Neisseria meningitides* (Sarangi et al., 2009) and *Streptococcus gordonii* (Azam and Shamim, 2014). Subtractive genomics implies the identification of pathogen conserved essential genes, which further acts as target for drug activity. It starts with removal of paralogous proteins of pathogen and the remaining proteins are subjected to non-homologs proteins subtraction within host and pathogen. Essential genes/proteins are then screened out of remaining genes and Database of Essential Genes (DEG) is used for this purpose as this database has all currently available genes which can sustain cellular life and potential targets for drug discovery (Zhang and Lin, 2009). Cytoplasmic proteins are preferred on other cellular localization areas as in cytoplasm drug acts in a superior way (Parvege, Rahman and Hossain, 2014). Current study uses the above explained strategy to screen out the potential drug target in human pathogen *V. cholerae* strains.

1.3.2 Potential Drug Target Selection

Target selection usually implies finding therapeutically significant agents (Knowles and Gromo, 2003) together with this, proper target identification suggests the relationship between drug and disease, which can further be analyzed for possible side-effects (Hughes et al., 2011). Druggability is important feature which should be retained while searching for potential targets and only those targets are of benefit, which pose both structural and functional features of druggability (Russ and Lampel, 2005). In the current study selected druggable target is 2, 3-dihydroxybenzoate-AMP ligase (EC: 2.7.7.58), which is an enzyme and gene name is vibE. 2, 3-dihydroxybenzoate-AMP ligase is an enzyme that catalyzes the chemical reaction where the two substrates of this enzyme are ATP and 2, 3-dihydroxybenzoate. This enzyme participates in siderophore biosynthesis. Siderophores have applications in medicine for antibiotics for improved drug targeting (Rusnak, Faraci and Walsh, 1989). Understanding the mechanistic pathways of these enzymes has led to opportunities for designing small-molecule inhibitors that block siderophore biosynthesis and

therefore bacterial growth and virulence in iron-limiting environments. Thus, vibE is used as a potential drug target.

1.3.3 Comparative Homology Modeling

Comparative homology modeling is a procedure which involves the prediction of structure (target) on comparison with aligned modeled structure (template) (Martí-Renom et al., 2000). As the sequence data is increasing the procedure of homology modeling is more in practice to generate protein structures but accuracy of modeled structures is mostly dependent on sequence alignment. By November 2015, available modeled protein structures in Protein Data Bank (PDB) (Berman et al., 2000) are 113816, which are far less from sequences available in Universal Protein Resource Knowledge Base (UniProtKB) (Boutet et al., 2007). Homology modeling provides the opportunity to fill this gap as this approach is cost-effective yet speedy as compared to experimental techniques and efficiency of this procedure can be increased by incorporating new *in silico* approaches (Cavasotto and Phatak, 2009). VibE lacks experimentally generated structure, so it is modeled via comparative modeling procedure and for that suitable target was accessed. For comparative analysis, different online servers for homology modeling were used, namely, SWISS-MODEL (Arnold et al., 2006), I-TASSER (Zhang, 2008), ModWeb (Pieper et al., 2004) and MODELLER version 9.14 (Eswar et al., 2008). Once models were retrieved from all web servers they were subjected to model validation and evaluation process to ensure that structure is in thermodynamically in a stable state (Garza-Fabre, Toscano-Pulido, and Rodriguez-Tello, 2012). To check the quality of modeled structure PDBSum (Laskowski, 2001), ERRAT value (Colovos and Yeates, 1993), ProSA (Wiederstein and Sippl, 2007), G-factor and bad contacts (Morris et al., 1992) were measured. These quality assessment measures are not only important to check model quality but also helped to improving drug target interactions which further increases the effectiveness of drug.

1.3.4 Molecular Docking

Molecular docking is a vital assistive tool in the process of computational drug design (Reddy et al., 2007). Its implementation in this field is evident from the innovative therapeutics procedures that have been developed recently, to counter diseases of various origins (Kumar, Chandra and Imran Siddiqi, 2014). Due to its well established role in CADD, the current work incorporates the molecular docking studies to investigate the inhibition mechanism of the target molecule. As an input to the docking procedure, both the active site of the target molecule and the inhibitors against it, were identified. These inhibitors were then docked in to the active site of the target protein using different docking procedures. Molecular docking essentially comprises of two major stages, firstly, the theoretical conformational space is explored to predict the spatial, physical and chemical complementarity between the ligand and the binding residues of protein to generate various conformations. Secondly, the strength of the interactions often denoted by either binding affinity or binding score is calculated (Meng et al., 2011). Various docking techniques however, may differ in terms of the algorithms used to achieve the aforementioned tasks.

1.3.5 Molecular Dynamics Simulation

Molecular dynamics simulations, provides the methodology for detailed microscopic modeling on molecular scale. Since, the nature of matter is to be found in the structure and function of its constituent building blocks. Molecular dynamics simulations are most flexible and extensively practiced computational techniques for studying the molecular structures and dynamic behavior of biological molecules. Molecular dynamics simulation studies can be a foundation for achieving some of the cutting edge results in the field of medicine and drug discovery by having an in depth knowledge of the proteins dynamic behavior involved in numerous diseases (Alonso, Bliznyuk and Gready, 2006). MD simulations play a promising role in the field of computer aided drug designing with improvements in both computational power and in algorithm design. Studying the details of microscopic events happening in only millionths of a second is impossible through experimental techniques, thus, MD simulations can replenish the details left in the experimental methods. Furthermore, MD simulations act as a bridge between the theoretical and experimental work (Allen and Tildesley, 1989).

1.3.5.1 Statistical Mechanics

Computational powers are not only complementing the experimental work but additionally assisting for future prognostics. In this context statistical mechanics is taking major part in the field of MD simulation. Statistical mechanics deals with study of a system at microscopic level including properties and the spontaneous fluctuations of individual atoms or molecules. Natural mechanism can be explored by using statistical mechanics. Thermodynamics properties like energy temperature, volume, and pressure can be allied with statistical mechanics to relate the microstates level to macroscopic level (Wereszczynski and McCammon, 2012). Ensemble is an important concept in statistical mechanics. Ensemble is stated as the assembly of all possible different microscopic states in a system is called ensemble. The four major ensembles in statistical mechanics are as follows:

1. **Canonical Ensemble (NVT)** is categorized by constant number of particle N, constant volume V and constant temperature T in thermodynamic studies.
2. **Micro canonical Ensemble (NVE)** is categorized by constant number of particles N, constant volume V and constant energy E.
3. **Isobaric-Isothermal Ensemble (NPT)** deals with constant particle number N, constant pressure P and constant temperature T value.
4. **Grand Canonical Ensemble (μVT)** is described as constant chemical potential μ, a constant volume V and a constant temperature T value.

1.3.5.2 Classical Mechanics

Classical mechanics deals with the study of bodies in motion following the general principles that were first given by Sir Isaac Newton. Classical mechanics is based on Newton's second law of motion which states that when a force "*F"* is applied on a particle of mass "*m"*, it will produce acceleration "*a"* in it. If force applied on each atom at molecular level is known then acceleration produced can be determined by integrating the second law of motion. Therefore it calculates the macroscopic properties of the system by defining velocities, positions and acceleration of a given system over the varying time scale (McCall, 2010). The equation of the second law of motion for the i^{th} particle is given by:

$$F_i = m_i a_i \qquad (1.1)$$

Where F_i is force applied on the particle, "m_i" is the mass of the particle "i" and "a_i" is the acceleration produced. The acceleration being second derivative of distance "r" and time "t" and substitution of these quantities give the following equation:

$$F_i = m_i \frac{d^2 r_i}{dt^2} \qquad (1.2)$$

When a force "F_i" is applied on a particle of mass "m_i" it can be described in terms of change of potential energy "v" to give the following:

$$F_i = -\frac{dv}{dr_i} \qquad (1.3)$$

So, "F_i" in equation 1.2 can be replaced by equation 1.3, and then they can be equated to give:

$$-\frac{dv}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \qquad (1.4)$$

During production run, the next coordinates for every particle in a system at any given time can be calculated using previous coordinates "r_o", velocity "v_o" and acceleration "a" at a given time "t".

$$\mathrm{a} = \frac{dv}{dt} \qquad (1.5)$$

Equation 1.5 denotes that for a particle in motion, its acceleration is dependent on the rate of change of velocity with respect to time.

1.3.5.3 Molecular Mechanics

Molecular mechanics is an extension of classical mechanics. It calculates the potential energy of the system of interest by applying force fields.

Force field of a molecular system can be given as the sum of individual energy terms:

$$E = E_{covalent} + E_{non\text{-}covalent} \quad (1.6)$$

$$E_{covalent} = E_{bond} + E_{angle} + E_{dihedral} \quad (1.7)$$

$$E_{non\text{-}covalent} = E_{electrostatic} + E_{van\ der\ waals} \quad (1.8)$$

1.4 Aims and Objective

The current study highlights and focuses on various computational approaches to explore the druggable genome of human pathogen *V. cholerae* strains, for identification of potential drug target. *In silico* subtractive genomics approach is applied to discover the complete genome of *Vibrio cholerae* three strains (O395, LMA3984-4 and IEC224) then a common novel drug target was identified, which will be effective against all three strains. *V. cholerae* is the most predominant strain involved in cholera and to date, not targeted for *in silico* drug designing approach. Resistance of *V. cholerae* against various antibiotics is increasing with time and thus the strain has become multi drug resistant. Consequently *V. cholerae* was selected for structure based drug designing process. Comparative modeling approach generated the structural coordinates of target, which was common to the three strains. Molecular docking and MD simulation highlighted the binding of different putative inhibitors. This study aims at finding the potential drug targets along with important inhibitors against cholera and the simulation study explores the dynamics of the docked target. This further aids in exploring the interactions and mechanism of the docked system. This inhibitor binding process will assist in synthesis of more powerful inhibitors against the pathogen which will consequently deal with resistance evolving ability of *V. cholerae*.

2 Methodology

Various *in silico* procedures accompanied by computational tools executed in current study to identify new drug targets that are stated in the following section.

2.1 System Specification

The whole strategy was divided into four steps; *in silico* subtractive genomics approach, homology modeling, molecular docking analysis and molecular dynamics simulation. Complete work was carried out using Intel (R) Core(TM) 2 Duo CPU E8600 @ 3.33 GHZ and the operating system used was Linux openSUSE 11.4. High performance computer cluster was used for carrying out the production runs (Figure 2.1). This operation facility was provided by Computational Biology Lab at National Centre for Bioinformatics Quaid-I-Azam University Islamabad, Pakistan. Furthermore, for extension of the simulations, assistance was acquired from Center for Chemistry and Biomedicine (CCB), University of Innsbruck, Austria.

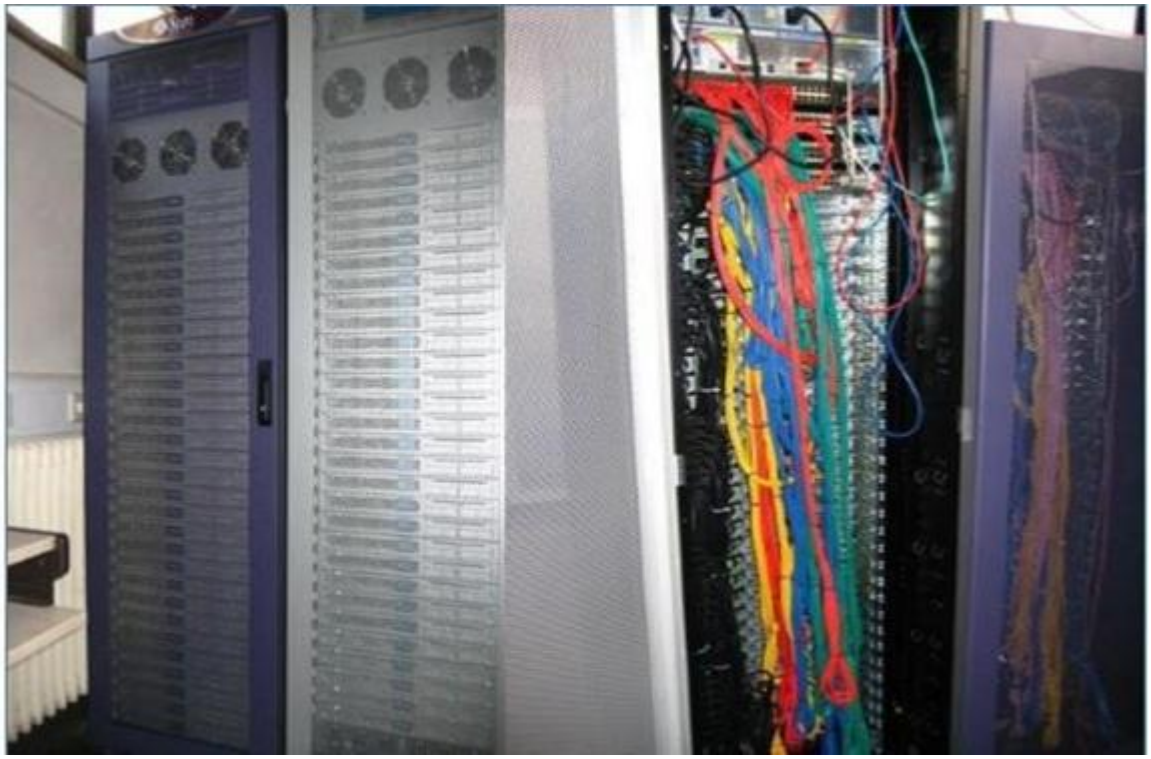

Figure 2. 1. *Cluster system used for computational studies including docking and simulation.*

2.2 Applied Computational Approaches

In silico approaches have been applied in current study to propose drug designing pattern against the human pathogen *V. cholerae*. The complete steps of the current work is summarized in Figure. 2.2. First phase involves the identification of potential druggable targets against *V. cholerae.* Complete protein dataset of the pathogen is subjected to numerous subtractive filters leading to identification of pathogen vital target followed by homology model of selected target. Docking protocols are employed to identify best binding inhibitor against the target and their important information about the interactions. In the last step, MD simulation assists in analyzing time dependent behavior and structural dynamics of the target.

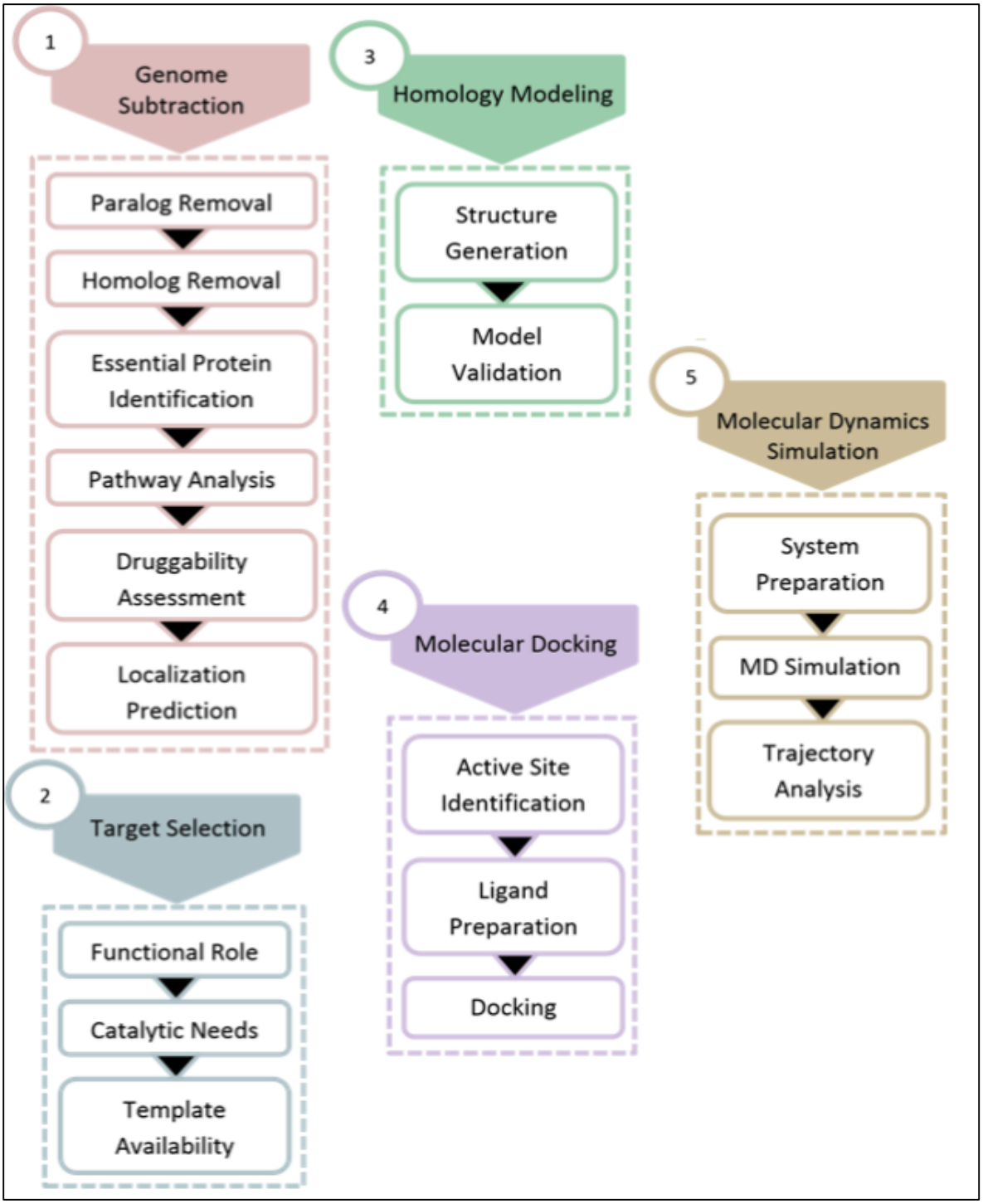

Figure 2. 2. *Schematic view of adopted methodology, highlighting major steps employed in study.*

2.2.1 Genome Subtraction

The first step in subtractive genomics was the availability of complete pathogen genome. UniProtKB *(http://www.uniprot.org)* (Boutet et al., 2007) was accessed to retrieve the complete genome set of *V. cholerae*. Once the genome was available, analysis was initiated, with the first step being removal of paralogous sequences.

2.2.1.1 Removal of Paralogous Sequences

Paralogous sequences were removed using CD-HIT (Li, Jaroszewski and Godzik, 2001) at 60% threshold in order to remove the redundancy in the data set with the proteome of all the three strains namely O395, LMA3984-4and IEC224 of *V. cholerae* in FASTA format, separately given as input.

2.2.1.2 Removal of Homologous Sequences

Homologous sequences were eliminated using Perl script which was provided at Computational Biology Lab of National Centre for Bioinformatics, Quaid-i-Azam University Islamabad, Pakistan. To attain pathogen specific drug target, resultant sequence set is subjected to BLASTp (Altschul et al., 1990) against human proteome (TaxID: 9606) at the expectation value (E-value) cutoff of 10^{-4}. This will retrieve *V. cholerae* proteins showing no significant similarity with human proteome. Non-homologous sequence selection guarantees the elimination of the threat of cross-reactivity in host.

2.2.1.3 Essential Proteins Identification

The Database of Essential Genes (DEG) holds 10,618 essential genes imperative for the survival and existence of eukaryotic and prokaryotic organisms (Zhang and Lin, 2009). Identified non-homologous protein sequences were then subjected to BLASTp search against DEG using E-value $10^{-10,}$ bit score = 100 and sequence identity of >= 30%. All these parameters were set in Perl script and essential genes for three strains O395, LMA3984-4and IEC224 were obtained.

2.2.1.4 Metabolic Pathway Analysis

Targeting proteins involved in pathogen-specific pathways, prevent any kind of harm to the host. Comparative pathway analysis was performed for the human and pathogen genome using KEGG Automatic Annotation Server (KAAS) available on: *http://www.genome.jp/kegg/kaas/*, an analytical tool supported by the Kyoto Encyclopedia of Genes and Genomes (KEGG) database *(http://www.genome.jp/kegg/).* KEGG is well known for its multiplicity in data collections and helps in predicting pathways (Kanehisa and Goto, 2000). KAAS in specific is a tool that assigns KEGG orthology (KO) numbers to the proteins by applying BLASTp to search against the KEGG GENES repository. Based on KO numbers, metabolic pathways are assigned to the studied set of sequences (Moriya et al., 2007). The essential non homologous proteins of *V. cholerae* obtained after search against DEG, were submitted to KAAS using *V. cholerae* genome as reference data set to increase specificity of results. Lastly, the predicted pathogenic pathways were manually matched to the human pathways to categorize them into two: common and unique. Common pathways were those prevailing in both the *V. cholerae* and the host. Whereas, unique, were exclusive to bacteria and therefore, focus of current research.

2.2.1.5 Druggability Assessment

Druggability screen was employed as a filtering criterion to the pathogen specific proteins. For this purpose DrugBank, an online tool which encloses information about experimental, investigational and approved drugs was used. (Knox et al., 2011). DrugBank version 4.2 (*http://www.drugbank.ca/search/sequence*), was used with default parameters to evaluate the druggability potential of non-homologous essential proteins. Only those proteins were selected which had a bit score > 100 and were not present in the category of withdrawn or illicit.

2.2.1.6 Localization Prediction

Genome subtraction final step was the prediction of the subcellular location of the screened, novel and essential protein targets. PSORTb3.0.2*(http://www.psort.org/psortb/)* (Nancy et al., 2010) and subCELlular LOcalization predictor (CELLO) version 2.5

(http://cello.life.nctu.edu.tw/) were used to accomplish this step. PSORTb, maintains a database of localization information on a wide range of bacterial and archeal species. It applies Support Vector Machines (SVMs), a machine learning technique that mines the curated dataset to predict the localization through the use of suffix tree algorithm (Cai et al., 2003). PSORTb is highly precise, analytical module which utilizes novel discoveries and technique in protein sorting (Nancy et al., 2010). Cytoplasmic proteins are selected as they have potential of becoming possible drug targets whereas surface membrane proteins are normally vaccine targets.

2.2.2 Drug Target Selection

Druggable essential proteins which were contributing in unique and essential metabolic pathways amongst different strains of *V. cholerae* were identified and the ones that were common in all three strains were used for analysis. VibE playing a significant role in the biosynthesis of siderophore pathway was selected for further *in silico* analysis. VibE sequence from all three strains were aligned using T-Coffee Alignment (Notredame, Higgins and Heringa, 2000) and showed 100% sequence identity amongst all. Therefore, sequence of vibE from *V. cholerae* were further explored for structural point of view. Due to unavailability of its structure, it was assisted via homology modeling and comparison of this sequence was done against PDB, by using BLASTp functionality supported by NCBI. Structural templates that showed at least 30% identity with > 90% query coverage were considered acceptable.

2.2.3 Comparative Homology Modeling

Since experimental structure was unavailable for vibE, comparative model building was carried out. At sequence level, *V. cholerae* 0395 vibE protein showed 97% coverage and 53% identity to the template structure PDB ID: 1MDB. Using the template as a guide, structural models were generated through MODELLER9.14 and a variety of web servers. A comprehensive comparison of the stereochemical properties was subsequently carried out to select the best modeled structure. In addition to MODELLER9.14, structure for vibE, was also obtained for comparative purposes through three web servers: SWISS-MODEL

(Schwede et al., 2003), ModWeb (Pieper et al., 2004) and I-TASSER (Wu, Skolnick and Zhang, 2007).

2.2.3.1 MODELLER

MODELLER is used to predict three-dimensional (3D) structures of proteins via homology or comparative modeling. Alignments of sequence with known related structures are provided as input and MODELLER generates a model.

2.2.3.2 SWISSMODEL

SWISSMODEL is an automatic server used to model tertiary and quaternary structure of protein. It can be accessed through ExPASY web server and Swiss PDB-Viewer *(http://swissmodel.expasy.org/).*

2.2.3.3 I-TASSER

I-TASSER combines repetitive use of Monte Carlo simulations with machine learning technique of neural networks and profile based alignment algorithm for detailed structural calculations. It can be accessed from: *http://zhanglab.ccmb.med.umich.edu/I-TASSER/.*

2.2.3.4 MODWEB

Incorporates use of MODELLER for model construction while integrating Position Specific Iterative-BLAST (PSI-BLAST) at the template selection step. It can be retrieved from: *https://modbase.compbio.ucsf.edu/scgi/modweb.cgi*.

2.2.3.5 Structure Evaluation

Discrete optimized protein energy (DOPE) score was used as a parameter to select the best model. Highly precise tools are accessed at National Institute of Health (NIH) server that provides all major structure validation tools through Structural Analysis and Verification Server (SAVeS) (*http://nihserver.mbi.ucla.edu/SAVES/).* The tools include: PROCHECK (Laskowski, Moss and Thornton, 1993), Errat (Colovos and Yeates, 1993) and ProSA-web (Wiederstein and Sippl, 2007).

2.2.3.6 PROCHECK

PROCHECK approximates stereochemical properties of protein model, including Ramachandran plot, G-Factor and Bad Contacts. Ramachandran plot represents the distribution of individual protein residues within the predefined allowed and disallowed regions derived from evaluation of phi and psi angles of experimental structures (Ramakrishnan and Ramachandran, 1965). G-Factor and Bad Contacts are measures of the main chain reliability reflecting the relative positioning of non-bonded atom relative to each other (Morris et al., 1992).

2.2.3.7 Errat

Errat is used for evaluating and refining the protein model. This verification algorithm works by statistically inspecting the non-bonded interactions among different atom types.

2.2.3.8 ProSA-web

ProSA-web, a tool used for validation and quality of protein structure on the basis of z-score. Physicochemical properties of the protein were studied using Expasy ProtPram server (Gasteiger et al., 2003).

2.2.4 Energy Minimization

Amongst the evaluated models, the best structure was selected and energy optimization of the protein model was carried out to improve its quality. The energy minimization procedure was performed on UCSF Chimera (Pettersen et al., 2004). Gasteiger charges were assigned to the protein and structural relaxation was achieved by application of 1500 rounds of minimization runs (750 steepest descent followed by 750 conjugate gradient) with a step size of 0.02 Å, under ff03.r1 force field.

2.2.5 Molecular Docking Protocol

The first step of molecular docking was the active site determination. Sequences from target protein and template were aligned and analyzed to check the conservation of residues. Putative active site residues were studied in detail with reference to the template structure. Secondly, potential inhibitors with reported activities against vibE collected from

BRENDA (Schomburg, Chang and Schomburg, 2002) (www.brenda-enzymes.org) were studied. A total 106 ligand library was constructed. 2D structures were drawn for all inhibitors via ChemDraw ultra of ChemOffice 2004 (Li et al., 2004). The minimization was carried out using MM2 force field of Chem3D Pro present in the same package. Docking process was carried out using minimized protein along with minimized ligand molecules. For docking Genetic Optimization for Ligand Docking (GOLD) (Jones et al., 1997) and AutoDock Vina (Trott and Olson, 2010) were used and GoldScore and binding affinities were calculated from them, respectively. On the basis of GoldScore best docked ligands were characterized. For docking results visualization and study of different interactions LIGPLOT (Wallace, Laskowski and Thornton, 1995), Visual Molecular Dynamics (VMD) (Humphrey, Dalke and Schulten, 1996), UCSF Chimera (Pettersen et al., 2004) and Discovery Studio (DS) Visualizer 3.5 (Visualizer, 2012) were used.

2.2.5.1 Docking Via GOLD

GOLD is a molecular docking program that performs automated protein-ligand docking using genetic algorithm (GA) (Jones et al., 1997). It offers several fitness functions including GoldScore. Current study incorporates GoldScore fitness function which predicts ligand binding modes by taking into account ligand torsional strain energy, receptor-ligand hydrogen bond energy, receptor-ligand van der Waals energy and ligand internal van der Waals energy. In the present study, Goldscore binding modes were ranked on the basis of molecular-mechanics based function:

$$\text{GOLD Fitness} = S_{hb\_ext} + S_{vdw\_ext} + S_{hb\_int} + S_{vdw\_int} + S_{tor} \quad \textbf{(2.1)}$$

In the above equation

- S_{hb_ext}signify: protein ligand hydrogen bond score
- S_{vdw_ext}signify: protein ligand van der Waals score
- S_{hb_int} signify : fitness because of intramolecular hydrogen bond
- S_{int} signify : the intramolecular strain in the ligand

Molecular docking was performed using default parameters in addition to usage of GoldScore. The population size representing the possible geometrical orientations of ligand was

100, selection pressure depicting the probabilistic ratio of a high fitness geometric coordinate to be selected as starting point for further orientation change was 1.1, number of genetic operations employed during a genetic algorithm execution was 10,000, number of islands was 1 representing a single population, niche size of 2 was used to maintain geometric diversity and operator weights determining the characteristic features of a GA run: migrate, mutate and crossover were 0, 100 and 100, respectively. Number of dockings is 10, and hydrogen atoms were added in the protein model. Docking results were analyzed and based on GoldScore fitness, the best hits were selected for further analysis.

2.2.5.2 Docking Via AutoDock Vina

Usage of AutoDock Vina led to the calculation of binding affinity of ligands. The distinctive feature of AutoDock Vina is its optimization framework which is inclusive of a Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm for local optimization. This combines both the scoring function and its derivative terms to expedite the optimization process (Trott and Olson, 2010). The AutoDock Vina protocol adopted in current study included preparation of ligand and protein files in pdbqt format followed by setting of docking grid to a size of 30 x 30 x 30 Å in the x, y, z axes, at -1.408, 25.029, 35.761 respectively. The entire docking procedure concluded with selection of the best docked chemical compounds which were analyzed using LIGPLOT (Wallace, Laskowski and Thornton, 1995) and ligand interaction mode of Molecular Operating Environment (MOE) (Chemical computing Group I, 2013) to understand the interactions that contributed to binding with the ligand

2.2.6 Molecular Dynamics Simulation

Molecular dynamics simulation study was conducted to study dynamic behavior of docked proteins. Assisted Model Building with Energy Refinement (AMBER) program was used for this purpose and analysis was performed using its different modules (Weiner and Kollman, 1981). Simulation of biomolecules performed in four steps which are given in Figure 2.3.

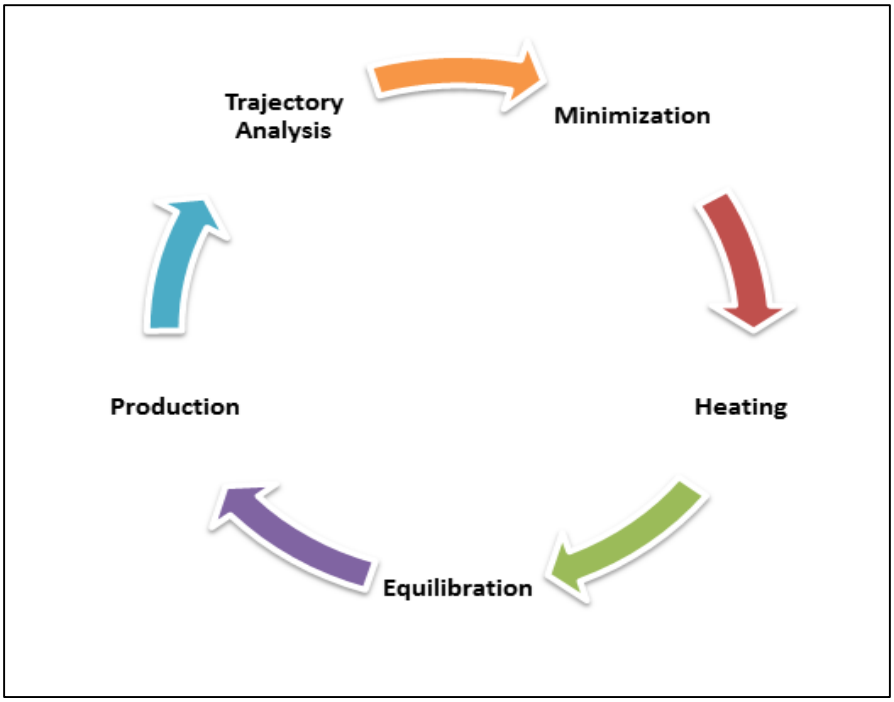

Figure 2. 3. *Steps involved in molecular dynamics simulations.*

2.2.6.1 System Preparation

MD simulations were performed using SANDER (Simulated Annealing with NMRDerived Energy Restraints) module in AMBER (Assisted Model Building with Energy Refinement) 10.0 suite of molecular dynamics program with the ff03.r1, GAFF (Duan et al., 2003), ff99SB forcefield (Pearlman et al., 1995). MD studies were carried out to investigate conformations of protein receptor, compute accurate energies and optimize the structures of docked complexes. AMBER force field (GAFF) using Antechamber program was used to generate force field parameters for the ligand molecule. Each complex was immersed in a cubic box of TIP3P water molecules with a 10 Å solute-wall distance (Figure 2.4). Net charge of -6 on vibE protein was neutralized by the addition of six Na+ ions. The energy of the solvated system was minimized before undergoing MD simulations.

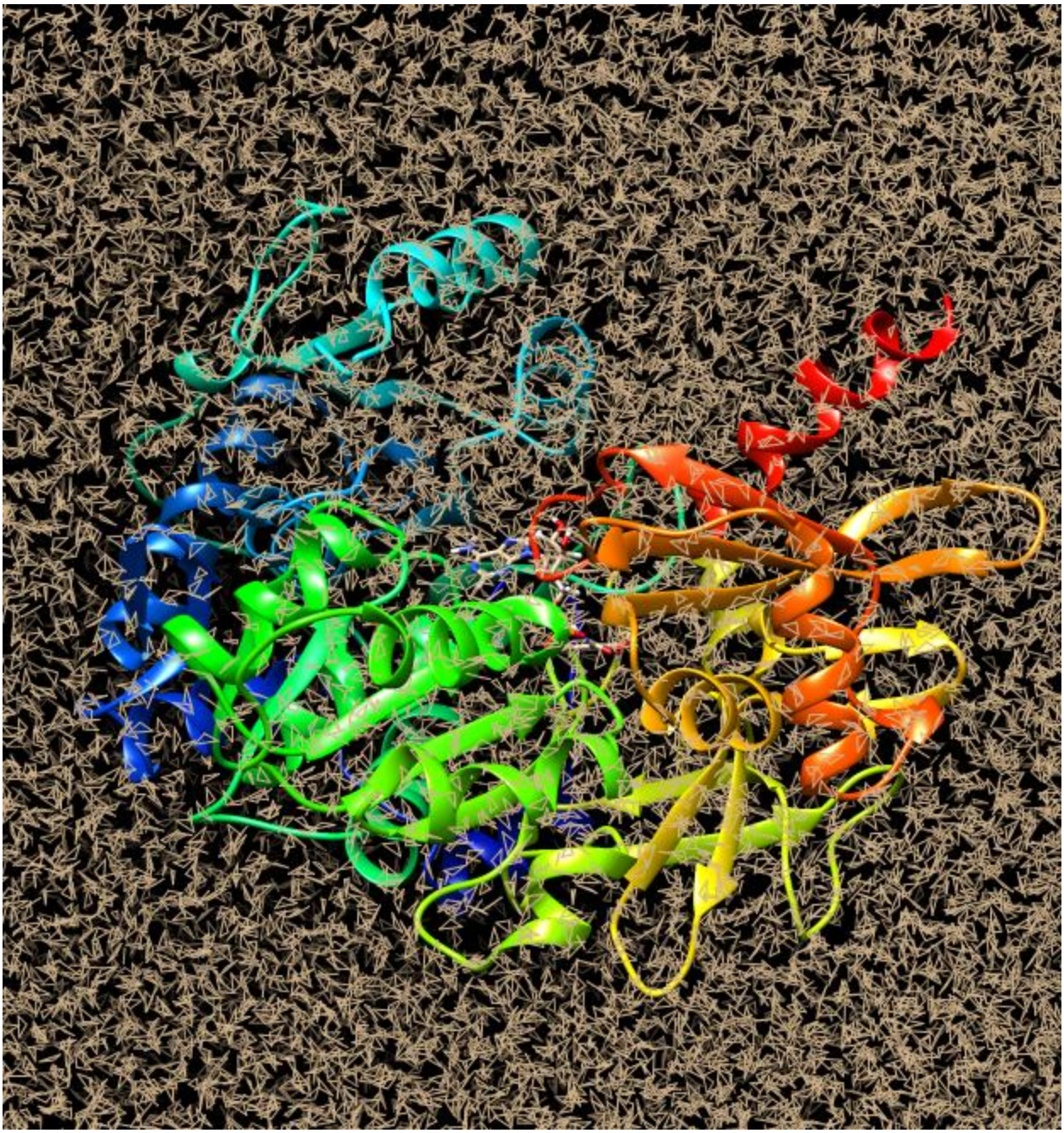

Figure 2. 4. *Solvation box surrounding docked protein.*

2.2.6.2 Minimization, Heating, Equilibration and Production

Following the system preparation minimization of proteins was performed to remove structural constraints. A total of 5000 steps, were split between 2500 steps of steepest descent and 2500 steps of conjugate gradient were performed with cutoff value of 8.0 Å. For heating of the system Langevin dynamics algorithm (Feller et al., 1995) was applied and heating was performed for 10 picoseconds. Along with this constant volume, constant temperature (300 K) and constant pressure (1 atm) was maintained. Equilibration was performed after heating for 100 picoseconds by keeping system temperature, volume and pressure constant. Production run was performed at the end which use SHAKE algorithm (Ryckaert, Ciccotti and Berendsen, 1977) for bond constraints. For production SANDER (Simulated Annealing with NMR Derived Energy Restraints) module was used and 70 ns for docked protein simulation run was performed. For whole simulation process canonical ensemble was used.

2.2.6.3 Simulation Trajectory Analysis

For analysis of results PTRAJ (Process TRAJectory) module of AMBER12 was utilized to make output files. Following four properties were calculated using PTRAJ and graphical representation were viewed in xmgrace (Vaught, 1996).

i) Root mean square deviation (RMSD)
ii) Root mean square fluctuation (RMSF)
iii) Radius of gyration (Rg)
iv) Beta factor (B-Factor)

2.2.6.3.1 Root Mean Square Deviation

Root mean square deviation (RMSD) is frequently used to compute the difference between the actually observed values and values predicted by a model. It gives the deviation of the coordinate of given set of atom in a time interval. Conformational changes at various time intervals and different folding procedures can be explored using RMSD values.

$$RMSD = \sqrt{\frac{1}{N} \sum_i di^2} \quad \textbf{(2.2)}$$

Where *N* represents number of compared atoms in a system and d_i represents the square root distance.

2.2.6.3.2 Root Mean Square Fluctuation

Root Mean Square Fluctuation (RMSF) is defined as the root mean square of averaged distance between the positions of atom from its mean position. RMSF is helpful in illustrating local fluctuations along the protein chain.

$$RMSF = \sqrt{\frac{\sum_{t_k} T(xi(t_k) - x}{T}} \quad \textbf{(2.3)}$$

Here "*T"* is signifying the time interval, "*xi"*: the position of an atom at a particular time and "*x":* the averaged position of the atom.

2.2.6.3.3 Beta Factor

Beta factor helps in analysis of dynamic property. It measures all the changes arising in the system due local vibrational and thermal movements. Beta factor is measured in terms of RMSF shown in equation below:

$$\beta \text{ Factor} = RMSF^2 \left(\frac{8\pi^2}{3}\right) \quad \textbf{(2.4)}$$

2.2.6.3.4 Radius of Gyration

Radius of gyration measures the compactness of the system. It can be calculated by the following equation:

$$Radius\ of\ gyration = \frac{\sum_{i=1} N\ m_i\ (r_i - r_{cm})^2}{\sum_{i=1} N\ m_i} \quad \textbf{(2.5)}$$

Here "*N"* is the total number of atoms, "m_i" shows mass and "r_i" represen*t* position vector of atom"*i"* and "r_{cm}" is the center of mass of molecules under consideration.

3 Results

3.1 Subtractive Genomic Approach

The complete number of proteins obtained at each step of subtractive genomics approach is shown in Figure 3.1.

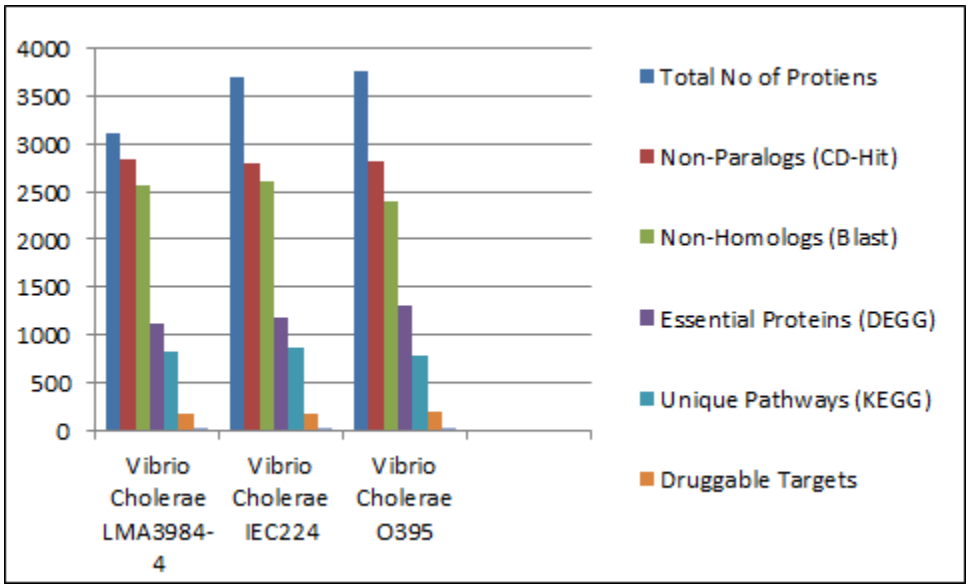

Figure 3. 1. *Overview of screened proteins obtained at the end of each subtractive genomic steps.*

3.1.1 Genome Retrieval

Current subtractive genomic study was performed on *V. cholerae's* strains namely, LMA3984-4, IEC224 and O395.These are completely sequenced strains and their genome was retrieved from UniProtKb. IEC224 and O395 are more evolutionary closer to each other than LMA3984-4 and their genomic features are mentioned in Table 1.

3.1.2 Non-Paralogous and Non-Homologous Proteins

After genome retrieval, first step i.e. non-paralog proteins removal was performed. Redundant protein sequences from the *Vibrio's* strains LMA3984-4, IEC224 and O395 were removed via CD-HIT program at 60% identity leaving 2994, 2981, and 2973 non-paralogous protein sequences in three strains, respectively. Subsequently, non-paralogous protein sequences were then subjected to Perl script for BLASTp to eliminate homologous protein sequences to host genome. Remaining sequences left as a result of BLASTp for all three strains were 1974 in LMA3984-4, 1987 in O395 and 1896 proteins in IEC224.

3.1.3 Pathogen Essential Proteins

For the next step again Perl script was used in which sequences were subjected to BLASTp against DEG database. In this way numbers were further reduced for all three strains. After this screening, essential genes identified were 1137 in LMA3984-4, 1212 in IEC224 and 1301 in O395, the remaining were non-essential proteins, which were not further included in analysis.

3.1.4 Metabolic Pathway Analysis

Metabolic pathways of the essential proteins were investigated through KEGG Automatic Annotation Sever (KASS) leading to the identification of potential drug targets involved in various crucial metabolic pathways of the pathogen. After accomplishing KEGG, number of sequences for LMA3984-4, IEC224 and O395 were reduced to 535, 520, and 501, respectively.

3.1.5 Druggability Assessment

In further step Drug Bank database was used to check the druggability potential of non-homologous essential proteins. This gives the essential drug target proteins for LMA3984-4: 74, for IEC224: 78, and for O395: 71.

3.1.6 Subcellular Localization

The retrieved sequences from all three strains were further analyzed using PSORTb version 3.0.2 to identify their subcellular locations. In LMA3984-4 strain, 9 were cytoplasmic,

similarly in IEC224 strain, cytoplasmic proteins were 8 and in O395; 9 cytoplasmic were observed. Cytoplasmic proteins were selected for identification of putative drug target as cytoplasmic proteins mostly consists of enzymes, which are important for bacterial growth.

3.2 Drug Target Selection

Unique pathways along with number of genes, which were identified in *V. cholerae* are listed in Figure 3.2. These pathways were present in all the three selected strains LMA3984-4, IEC224 and O395. Total 8 proteins were involved in these pathways which were common in all three strains. Out of these drug targets six proteins were short listed on various parameters outline in Table 3.1. In current study, 2, 3-dihydroxybenzoate-AMP ligase i.e. vibE, involved in biosynthesis of siderophore with EC number EC: 2.7.7.58 was successfully selected for CADD analysis. VibE is essential protein in all three strains and further study was carried out on this protein as it is an effectual drug target.

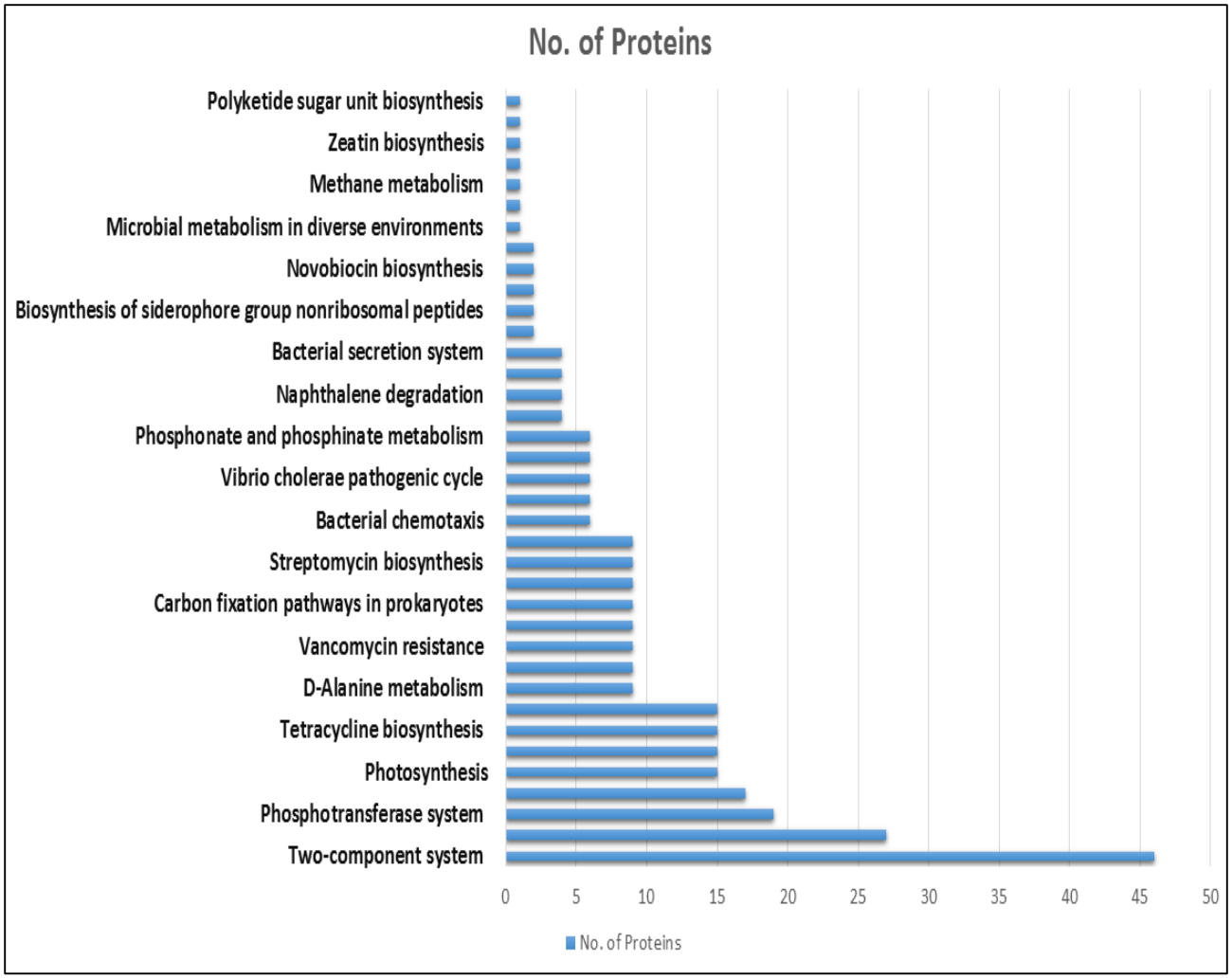

Figure 3. 2. *Number of proteins involved in the unique metabolic pathways of Vibrio cholerae.*

Table 3.1. *Features used to identify feasibility of targets for CADD analysis.*

Gene Name	Unique Metabolic Pathway	Protein Length (AA)	Structure Available	Query Coverage with Target	Sequence Identity with Target	Sub-Cellular Location
vibE	Biosynthesis of siderophore group	543	No	97%	53%	Cytoplasmic
PTS-EI.PTSI	Phosphotransferase system	573	No	99%	75%	Cytoplasmic
envZ	Two-component system	438	No	59%	48%	Cytoplasmic
glnG	Two-component system	466	No	79%	42%	Cytoplasmic
sig2	Vibrio Cholerae Pathogenic Cycle	335	No	67%	45%	Cytoplasmic
torD	Two-component system	220	No	91%	38%	Cytoplasmic

3.3 Comparative Molecular Modeling

Since experimental structure was unavailable for vibE, comparative model building was carried out. At sequence level, *V. cholerae* vibE protein showed 97% coverage and 53% identity to the template structure PDB ID: 1MDB. Using the template as a guide, structural models were generated through MODELLER9.14 and a variety of web servers. A thorough comparison of the stereochemical properties was subsequently carried out to select the best modeled structure (Table 3.2). Based on the quality assessment measures obtained for the various homology models, Model number 2 generated via MODELLER9.14 was selected for further processing. In addition to providing significant coverage, Model 2 showed strong stereochemistry with no residues in disallowed regions and no bad contacts (Table 3.2). Moreover, when superimposed, the backbone atoms showed RMSD of 0.237 Å, which is representative of high accuracy of the model, indicating high degree of structural similarity of the generated structure with the respective template thus indicating the accuracy of predicted model. To remove steric clashes the model building procedure was followed by energy minimization in order to relax the overall structure and allow adjustment of side chains. An additional benefit of optimization procedure was the improvement in the ERRAT quality factor which increased from 78.37 to 81.70. An illustrative view of the superimposed target-template structures is shown in Figure 3.3. The homology model served as a starting point for the docking and subsequent simulation procedures. Physicochemical properties of selected model 2 are also given in Table 3.3. Ramachandran plot of the selected model is shown in Figure 3.4 (a) where maximum residues are present in the most favored regions. Furthermore, the Z-score of the selected optimum model is plotted in Figure 3.4(b).

Table 3. 2. *Stereo-chemical properties of comparative homology modeled structure.*

Structure Resource	Number of Residues				G Factor	Bad Contacts	Z Score
	[A,B,L] Allowed region	[a,b,l,p] Additionally allowed region	[~a,~b,~l,~p] Generously allowed region	Disallowed region			
MODELLER (1)	93.5%	5.6%	0.4%	0.0%	0.4	0.2	-9.78
MODELLER (2)	**93.5%**	**6.9%**	**0.8%**	**0.0%**	0.8	**0**	**-9.84**
MODELLER (3)	92.3.4%	5.8%	0.4%	0.0%	0.6	0.1	-9.94
MODELLER (4)	93.%	6.1%	0.4%	0.0%	0.2	0.2	-9.97
MODELLER (5)	93.5%	6.1%	0.2%	0.2%	0.5	0.3	-9.86
I-TASSER	79.0%	17.2%	1.0%	1.7%	0.4	0.8	-10.6
ModWeb	90.3%	6.3%	0.4%	0.0%	0.7	0.5	-9.01
Swiss-Model	85.4%	8.5%	1.5%	0.6%	0.2	0.6	-10.72

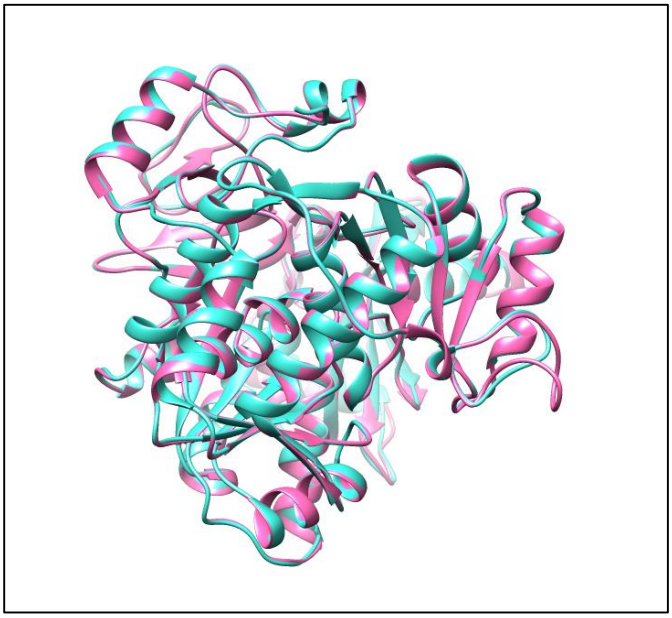

Figure 3. 3. *Superimposed structures of template IMDB (blue) and target vibE (pink).*

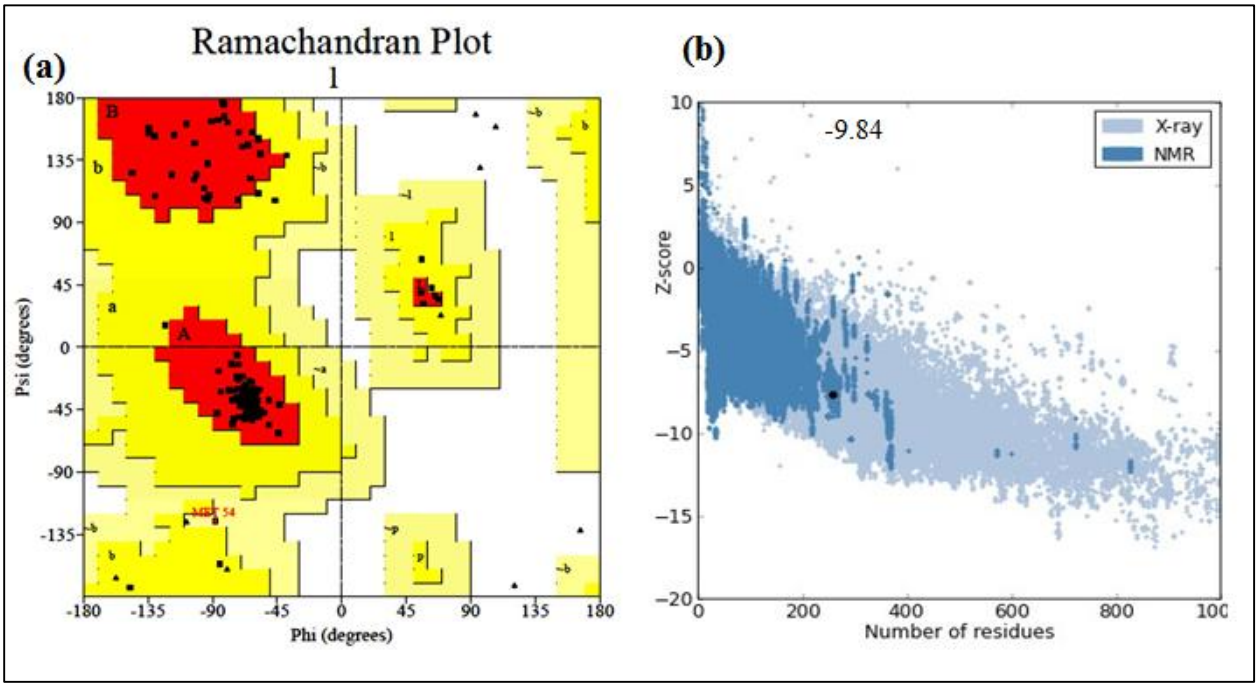

Figure 3. 4. *(a) Ramachandran plot of the selected model. (b) Z-score of the selected optimum model.*

Table 3. 3. *Physicochemical properties of vibE using ExPASy ProtParam tool.*

Physicochemical Properties	Values
Number of amino acids	543
Molecular weight	60123.4
Theoretical pI	6.02
Instability index	36.48
Aliphatic index	92.43
Grand average of hydropathicity (GRAVY)	-0.144
Total number of negatively charged residues (Asp + Glu)	55
Total number of positively charged residues (Arg + Lys)	47

3.4 Molecular Docking

To initiate the docking procedure active site information is a necessity. Binding patterns observed in active site are explained below.

3.4.1 Active Site Identification

The *V. cholerae* vibE structural orthologs were identified using BLASTp search against PDB. Sequences retrieved for the top scoring hits from three bacterial species served as input to ClustalO, which was used to build alignment. The reported work corresponding to the sequences used in MSA confirmed the conservation pattern observed at the sequence level. Since, the structures for these orthologs were obtained in inhibitor bound state, information about the topological features and binding pocket residues was available. The active site of protein was further confirmed via literature (May et al., 2002).

3.4.2 Inhibitors Selection

In current study inhibitors accessed from the BRaunschweig ENzyme Database (BRENDA) and the literature were employed for docking studies. Total 106 ligands were docked into the active site of target using GOLD and AutoDock Vina for calculation of GOLDScore and binding affinities, respectively. Along with this, preferred binding pocket orientation of active compounds was also identified. 2, 3-dihydroxybenzohydroxamoyl adenylate, was the top scoring compound in the active site. Docking results also depicted the hydrogen bonding between the potential inhibitor and active site residue.

3.4.3 Binding Analysis

GOLD was used to dock the prepared ligand molecules into the active site of the target. The resultant binding affinities were also calculated using AutoDock Vina. Results obtained from GoldScore values ranged from 35.31 to 75.7 with binding affinities between -4.7 and -7.8 kcal/mol. The highest score of 75.7 was achieved for compound 103, with binding affinity of -7.5 kcal/mol. For top 10 compounds the docking scores along with the respective binding affinities, arranged in descending order of GoldScore values are delivered in Table 3.4. Comprehensive visualization analysis conceded out through UCSF Chimera, LIGPLOT, DS Visualizer and MOE, exposing the conformational details along with the preferred orientation of the ligand binding. Ligand positioning within the active site is outlined in Figure 3.5.

Table 3. 4. *Docking results of inhibitors arranged in descending order of GOLDScore with corresponding binding affinities.*

Sr. No.	Compound No.	GoldScore	Binding Affinity (kcal/mol)
1.	103	75.7	-7.5
2.	76	70.4	-7.4
3.	106	70.2	-6.8
4.	81	69.5	-7.8
5.	96	69.4	-7.6
6.	79	58.0	-6.9
7.	105	57.7	-7.2
8.	63	56.8	-7.1
9.	20	55.9	-6.8
10.	15	55.3	-7.4

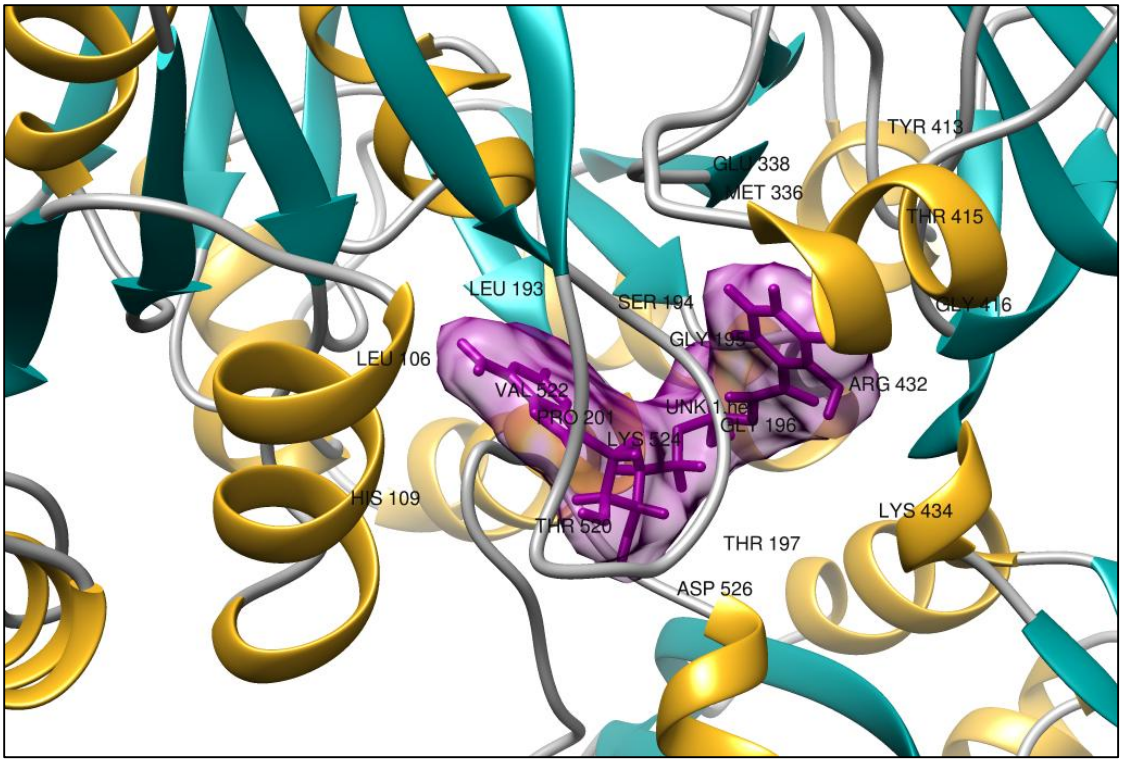

Figure 3. 5. *Best docked inhibitor (blue) in the active site of Vibrio cholerae vibE.*

The binding of compound 103 was observed at the domains of protein and residues involved in electrostatic interaction His 238, Lys 524, Leu 193, Gln 192, Gly 195, Lys 434, Gly 196, Aps 526, Thr 197 (Figure 3.6). Additionally, LIGPLOT image i.e. Figure 3.8 showed the presence of hydrogen bonds between ligand and target. Ligand oxygen moiety formed two hydrogen bonds with residue Glu 338 having 2.94 Å and 2.75 Å distance and His 238 atom developed hydrogen bonding with ligand at the distance of 2.81 Å. Moreover, Asp526 making a hydrogen bond of length 3.02 Å. Along with this hydrogen bond details of ligand with target residues are given in Table 3.5. The residues Pro 201, Val 522, Gly 416, Ile 525 along with other residues were involved in hydrophobic interactions (Figure 3.8).The MOE illustration further helped to visualize minor details e.g. acidic basic residues, side chain acceptor, side chain donor, ligand exposure and the receptor exposure (Figure 3.7).

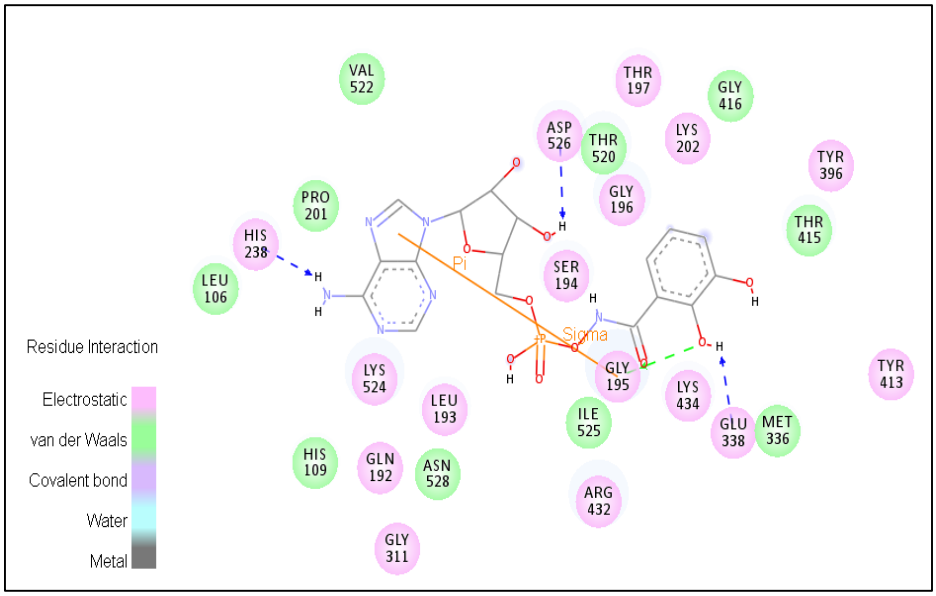

Figure 3. 6. *DS Visualizer 2D depiction of compound 103 interactions with the ligand.*

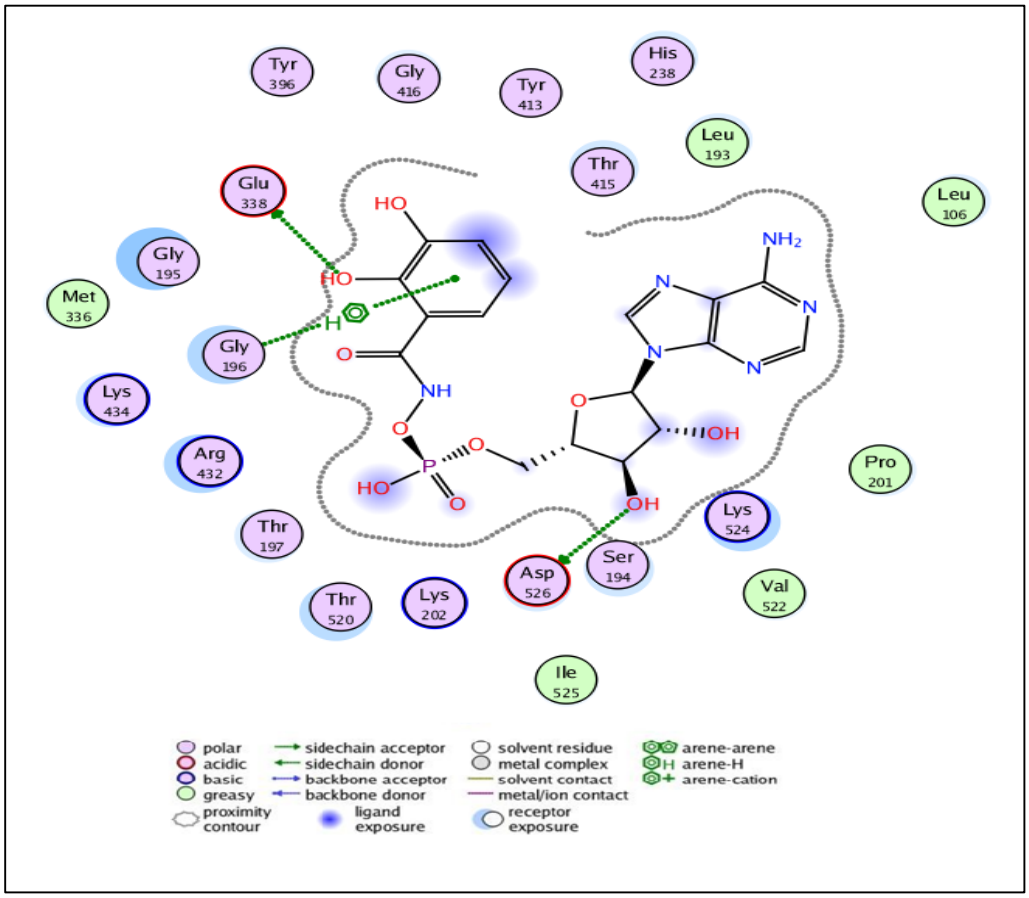

Figure 3. 7. *MOE ligand interaction image showing bonded and non-bonded interactions of inhibitor bound vibE.*

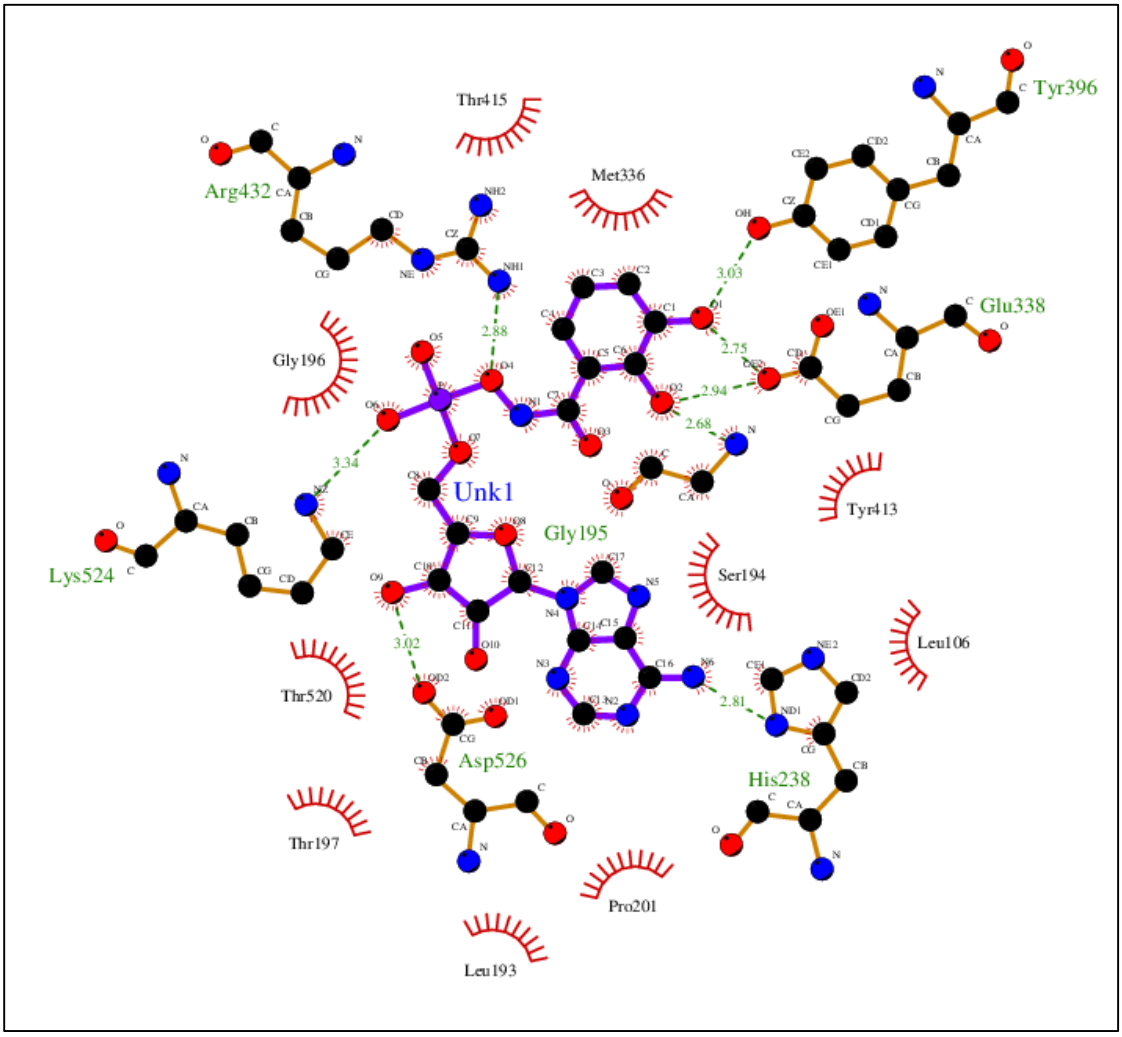

Figure 3. 8. *Interaction of ligand with vibE, highlighting interacting residues through LIGPLOT.*

Table 3. 5*. Hydrogen bond details of best docked compound with important interacting residues.*

Protein Interacting Atom	Ligand Interacting Atom	Distance (Å)
194 SER O	UNK N	2.98
194 SER HG	UNK N	3.89
194 SER O	UNK H	3.45
195 GLY O	UNK N	2.21
195 GLY HA2	UNK O	3.09
195 GLY HA3	UNK N	3.88
196 GLY HA2	UNK N	3.49
197 THR N	UNK H	2.57
238 HIS ND1	UNK N	2.81
413 TYR HH	UNK O	3.20
432 ARG HD3	UNK N	2.70
432 ARG HE	UNK N	2.88
432 ARG NE	UNK O	2.77
432 ARG HD3	UNK O	2.53
432 ARG NH1	UNK H	3.82
434 LYS HE2	UNK O	2.47
524 LYS NZ	UNK H	3.42
526 ASP OD2	UNK H	3.62

The best docked inhibitor from AutoDock Vina was again compound 103 with a binding affinity of -7.5 kcal/mol. The inhibitor interacting with the target molecule is shown in figure 3.9.

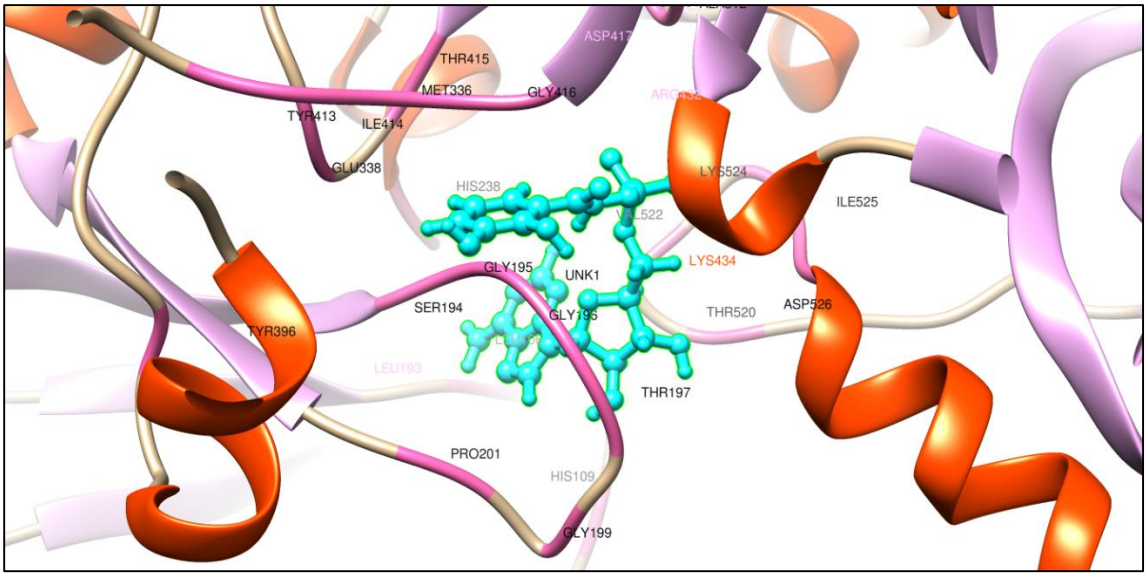

Figure 3. 9. *Best Vina docked inhibitor (blue) in the active site of Vibrio cholerae vibE.*

3.5 Molecular Dynamics Simulation

Molecular dynamics simulation provided a meaningful insight into the structural basis of vibE which is a potential candidate for drug target. Simulation of docked complex was carried out for 70 ns. For trajectory analysis PTRAJ module of AMBER was utilized. Physical properties including root mean square deviation (RMSD), root mean square fluctuation (RMSF), B-factor and radius of gyration of the system are taken into account along with conformational changes of vibE in the presence of inhibitor within the hydrated system are being studied.

3.5.1 Root Mean Square Deviations (RMSD)

RMSD graph for docked complex was unstable in first few nanoseconds. Following simulation time the docked complex become stable. The average RMSD value of 2.84 Å was

observed, reaching the maximum value of 3.81 Å. Overall, the pattern of RMSD graph supports slight shifts within the structural framework of the protein-ligand complex. Figure 3.10 represents the RMSD graph of the protein complex.

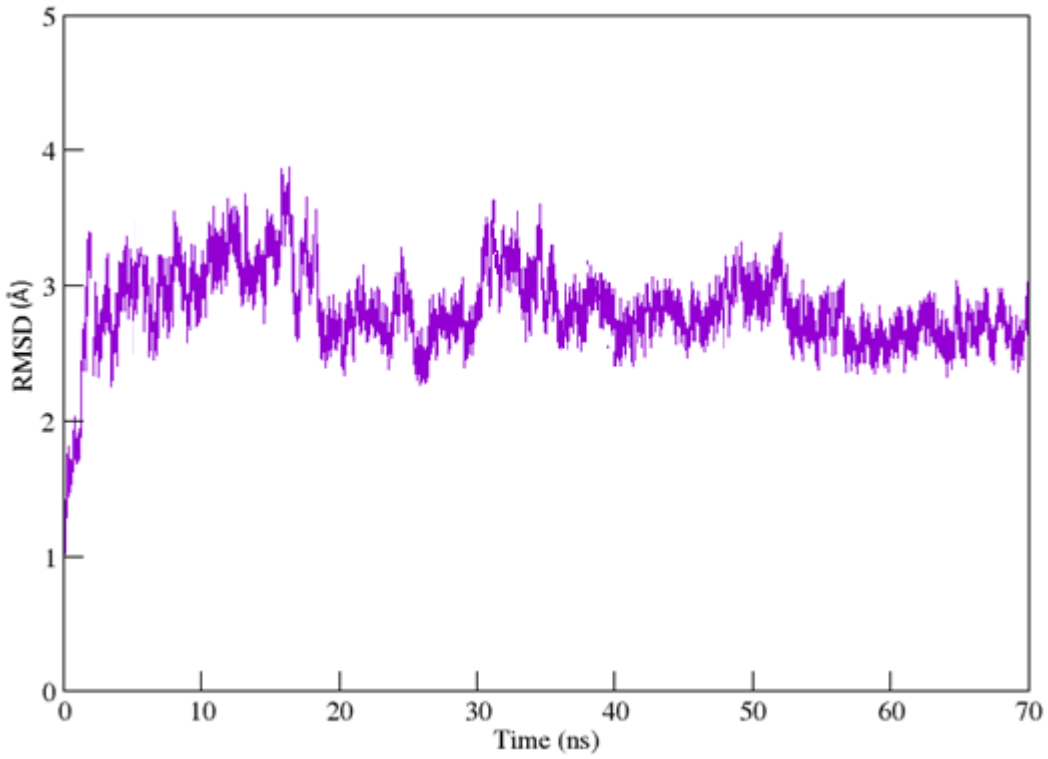

Figure 3. 10. *RMSD plot of docked vibE protein complex for 70 ns simulation run.*

During the simulation runs no major domain shifts were observed however, after the 20 ns helix (highlighted) was replaced by loop and at the end of 70 ns some secondary structure rearrangements were noticed (Figure 3.11). Ligand displacement over the 70 ns simulation run is shown in Figure 3.12.

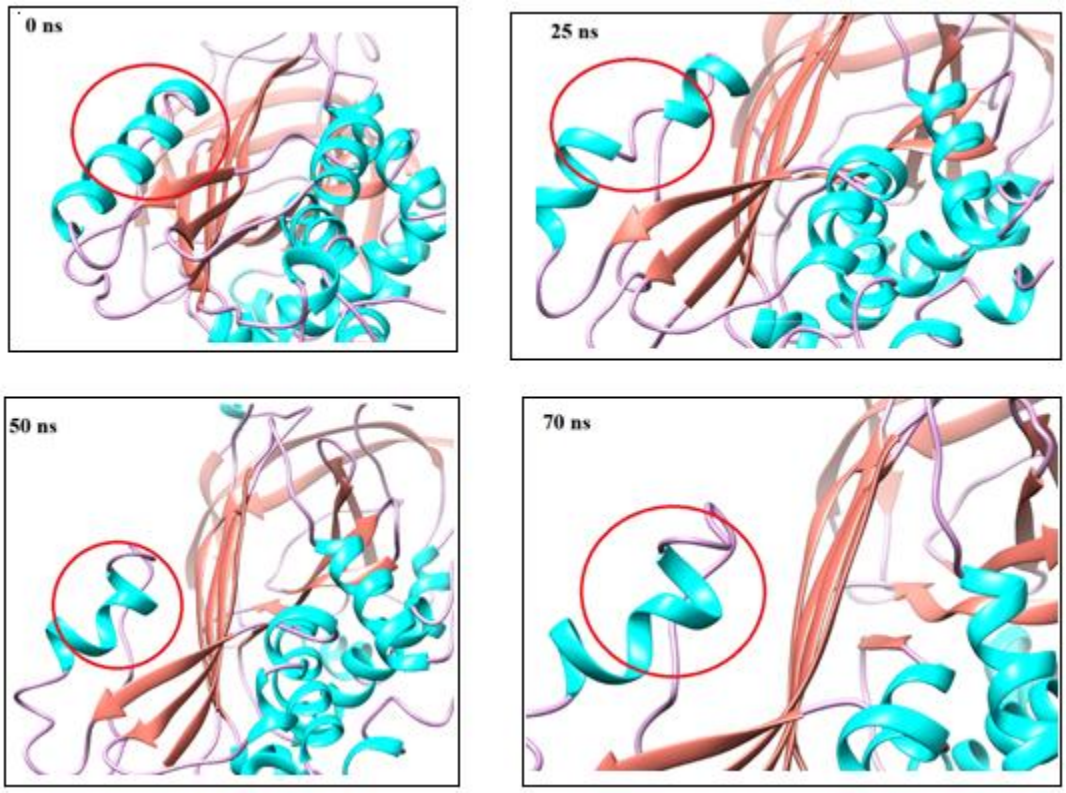

Figure 3. 11. *Snapshots taken of docked protein vibE at 0 ns, 25 ns, 50 ns and 70 ns timescale. The red circle highlights the changes that were observed.*

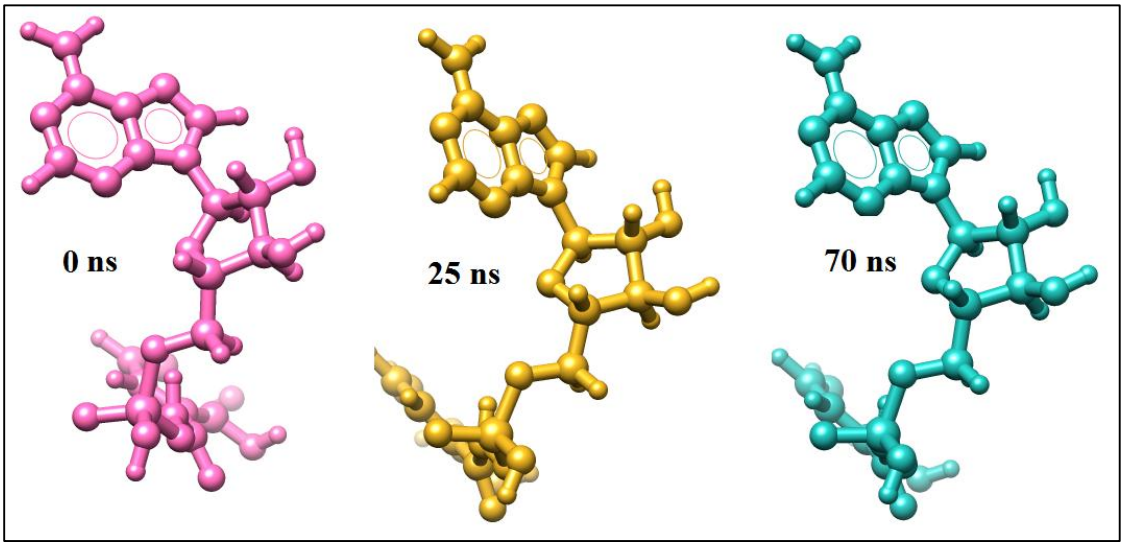

Figure 3. 12. *Ligand displacement of vibE docked complex at 0 ns, 25ns and 70ns.*

3.5.2 Root Mean Square Fluctuations (RMSF)

RMSF is investigated to have insights into structural dynamics of residues of the protein. It assisted to identify and understand the structurally flexible and rigid regions of the potential drug target. Average root mean square fluctuation for the docked system for 70 ns was 1.49 Å (Figure 3.13), while the maximum value is 5.82 Å (Figure 3.13). It is a common observation that the terminal regions show more fluctuations than any other part of protein. Therefore, RMSF trajectory analysis of normal binding site complex showed higher fluctuations for the amino acid residues present at N terminal region. In comparison to this active site residues revealed lesser fluctuations. From above discussed observation it can be interpreted that active site is more stable in a complex where as N terminal region is highly flexible. In comparison with the entire protein, higher fluctuations were observed at two regions: 62-72, 350-355 for the docked protein signifying disorderness of atoms in these regions.

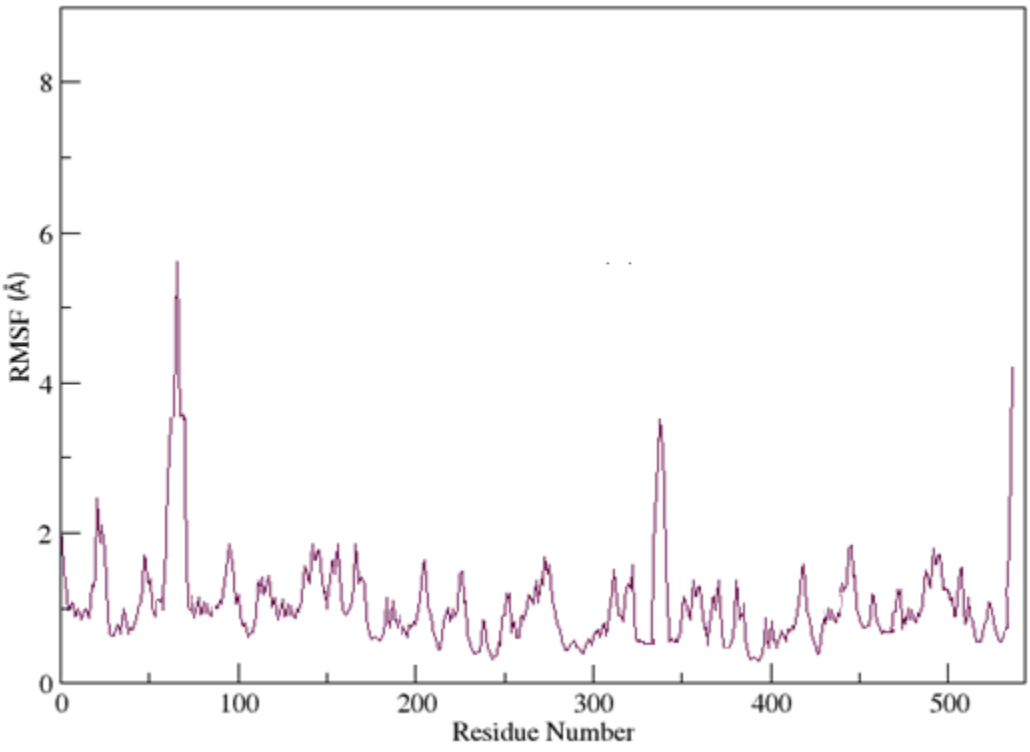

Figure 3. 13*. RMSF plot of docked vibE protein over 70 ns simulation run.*

3.5.3 β-Factor Analysis

Beta factor is the measure of thermal dependent disorderness and structural stability of system. Analysis exhibit beta factor not greater than 2055 Å^2 with the average beta factor of 77.1 Å^2 for ligand bound protein, presenting conformational homogeneity (Figure 3.14). B-factor value greater than 250 Å^2 indicates possible atom's disorderness. The graphs depicts overall thermal stability of the docked complex.

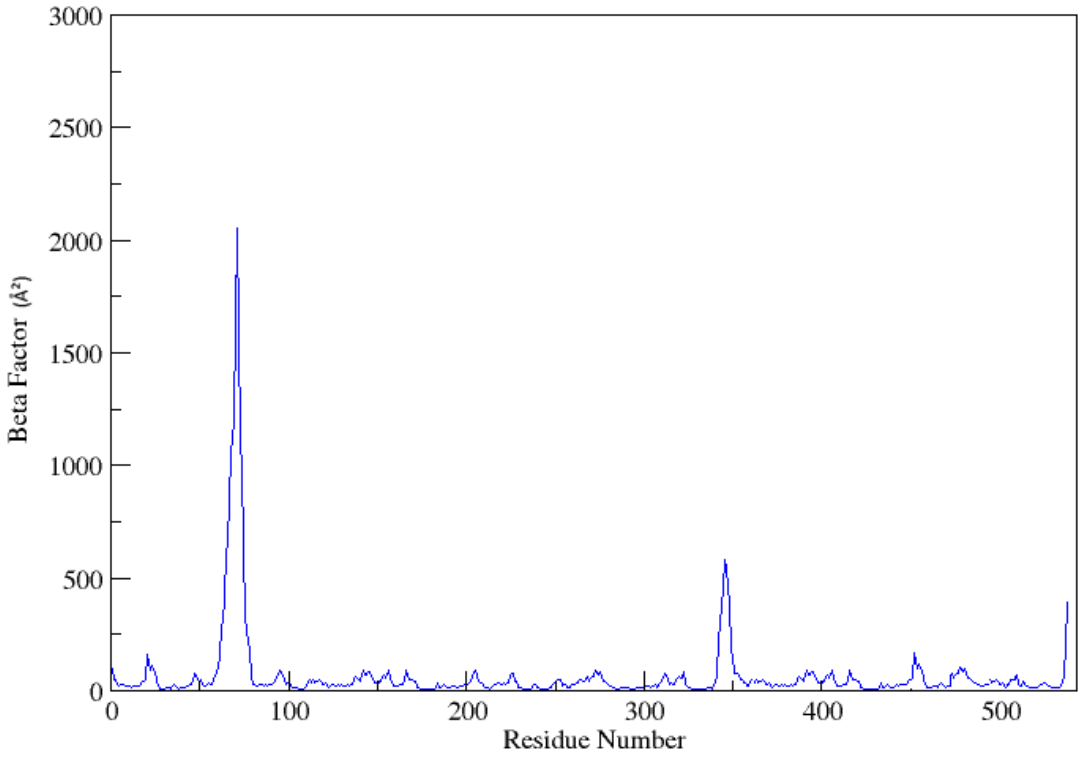

Figure 3. 14. *β-Factor graph of docked vibE protein over 70 ns simulation run*

.

3.5.4 Radius of Gyration

Compactness and stability of a protein structure is determined by its radius of gyration (R_g). Reduction of radius of gyration values specified the stability of the system. The average value of 22.74 Å observed for docked protein complex, with a maximum value of 23.23 Å,

denotes stability of the protein structure (Figure 3.15). This implies that docked complex exhibits stable and compact system according to the value of radius of gyration.

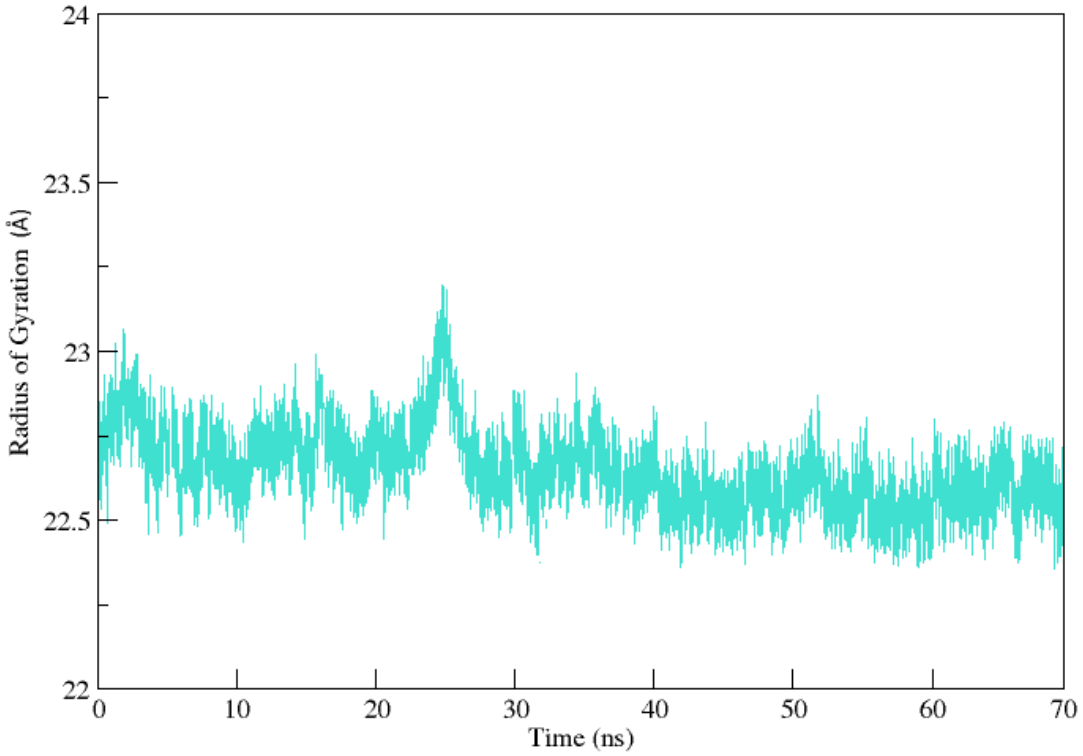

Figure 3. 15. *Radius of gyration of docked protein vibE over 70 ns simulation time period.*

4 Discussion

Pathogens responsible for different diseases are found everywhere in our environment, from hot springs to snow glaciers. The use of antibiotics is increasing day by day to encounter the attack of the pathogens. With the passage of time the pathogens are evolving and they are developing resistivity leading to multi drug resistant strains. Therefore, the treatment of infectious diseases is becoming difficult. Consequently, the demand of novel, specific and more effective therapeutic agents is increasing.

Cholera, targeted in present study, caused by *V. cholerae*, is used for identification of potential druggable candidates. *In silico* subtractive genomic approach, which is a therapeutic target screening technique, was applied to screen the pathogenic genome. The procedure initiated with the proteome analysis, the obtained potential drug candidate had two significant features: firstly, it was specific to the pathogen thus, limiting cross reactivity; secondly, it was crucial to the pathogen metabolism. O395, LMA3984-4 and IEC224 are *V. cholerae* strains which were used in the current study to combat cholera with the same drug. Total 3747, 3076 and 3677 proteins are present in three strains, respectively. For the drug molecule to bind specifically to the target CD-HIT tool was used and it removed all the duplicated sequences from pathogen genome. Additionally, no homologous sequences between the host and pathogen were guaranteed, and to achieve this, BLASTp was performed against RefSeq database, which filtered non-homologous pathogen sequences. Essential proteins of the pathogens, which are crucial for their growth in host are then subjected to metabolic pathways association predictions. In pathways analysis thirty five unique bacterial pathways were identified which could be targeted for therapeutic effect. All the aforementioned subtractive genomic steps were applied on the three strains independently resulting with 9 (O395), 8(IEC224) and 9 (LMA3984-4) potential cytoplasmic drug target proteins. Selected therapeutic target was vibE, which was subjected to the CADD procedure.

Comparative homology modeling was performed to model vibE, to identify the possible binding modes of the ligands. Template selected for modeling having the PDB ID: 1MDB showed 97% coverage and 53% sequence identity. The model was evaluated with different

tools like ERRAT, PDBSum and Verify3D, which tested the quality of structures generated. Model 2 was selected as it had minimal bad contacts 0 and highest G factor 0.8. MODELLER model was considered the most reliable for conducting docking and simulation studies. Among the web servers ModWeb generated the best model. RMSD is calculated to check the similarity between two proteins, which analyze the ensembles of backbone atoms for proper conformation (Maiorov and Crippen, 1994). RMSD value calculated for target and template was 0.237 Å, low value means generated model was of good quality and had similar main chain fold as that of template chain. In order to relax the overall structure and allow adjustment of side chains to remove steric clashes energy minimization was done. An additional advantage of energy minimization procedure was the improvement in the ERRAT quality factor which increased from 78.37 to 81.70. Comprehensive physicochemical properties of the modeled structure are mentioned Table 3.3.

GOLD and AutoDock Vina were used to dock ligand molecules into the active site of the target. Total 106 ligands were docked in active site present in vibE. Compound 103, was the best docking ligand and showed the highest GOLDScore 75.7 with binding affinity of -7.5 kcal/mol. Interaction studies of compound 103 with protein was visualized by LIGPLOT, DS visualizer and UCSF Chimera. Ligand oxygen moiety formed two hydrogen bonds with residue Glu 338 having 2.94 Å and 2.75 Å distance and His 238 developed hydrogen bonding with ligand at the distance of 2.81 Å, while DS Visualizer highlighted the hydrogen bonds with Asp526, Glu338 and His238. The residues Pro 201, Val 522, Gly 416, and Ile 525 along with other residues were involved in hydrophobic interactions. Moreover, π-interactions interactions were also present between ligand and target protein. The docking study of vibE, further needed an understanding of the structural adjustments made upon ligand introduction into the system. However, it provided this information within the context of a static environment. In order to insinuate the dynamic conduct, simulation protocol was carried out that provided eloquent insights into the structural basis of druggability potential of vibE. This was followed by trajectory analysis to assess various characteristics of the docked system. Molecular dynamics simulation allows to unveil the time dependent dynamic behavior of biomolecules. It provides the insights about the region of molecule which is involved in dynamic behavior of system (Azam, Uddin and Wadood,

2012). The application of simulation to *V. cholerae* vibE, specified structural stability of the system over the studied time scale.

Properties namely RMSD, RMSF, B-factor and Radius of gyration were plotted as a function of time to implicate the biomolecular arrangements within a solvated environment. The deviation of the backbone Cα atoms was studied for a time period of 70 ns. The inhibitor bound vibE with average RMSD value of 2.84 Å was observed over the studied time scale, shows structural stability. It also denotes that the ligand placement is well supplemented within the active site and does not disrupt the protein chemistry. Overall, the pattern of RMSD graph supports slight shifts within the structural framework of the protein-ligand complex. In accordance with the RMSD graph, conspicuous structural alterations occurred at a time of 10ns, 25ns and 50ns, after which the protein acquired a stable conformation. Figure 3.12 represents the structural and conformational alterations that occurred in the docked protein complex, at different time lapses. The graph indicates an alternate increasing and decreasing trend at the 20^{th}ns. This is due to the structural shift of an alpha helix into a turn at one location. Obvious structural changes occurred from 25^{th}ns to the 50^{th}ns. The protein structure however, maintains the alterations in its conformity, from the 50^{th}ns onwards, till the 70^{th}ns. Also, the structural compactness of the ligand-bound vibE, measured as radius of gyration (R_g) averaged at a value of 22.74 Å.

As a measure of atomic fluctuations, RMSF helped distinguish between the structurally flexible and rigid loci of the drug target. The average Cα fluctuation is 1.49 Å. The RMSF graph designates the residue locations, instigating the changes. The highest peak nominates 11 residues in total that form a loop, notably Val62, Thr63, Glu64, Asn65, Thr66, Thr67, Glu68, Lys69, Ala70, Ala71 and Thr72. Changes in this loop are the major source of the peaks in the RMSD graph. The other residues being altered are Leu350, Val351, Asn352, Ala353, Pro355 and Lys354. Amino acid residues that incur the fluctuations are away from the active site pocket of the protein. Thus no major changes occur at the active site. The pattern of beta factor for the docked protein complex is consistent with the RMSF trend. Regions having greater fluctuations as identified in the RMSF analysis exhibit average beta factor of 77.1 $Å^2$ for ligand bound protein, presenting conformational homogeneity.

In silico approach adopted in current study highlighted significant results at various stages of analysis. The structure dynamics of docked protein and its simulation analysis provided important insights, which can be practiced to increase the efficacy of drug against the *Vibrio* strains along with drug target specificity and selectivity and to cure infections caused by pathogen *V. Cholerae*.

Conclusion

The strategic direction applied for the scrutinizing therapeutic candidates in the emerging and evolving gram negative, multi drug resistant bacterial pathogen, *V. Cholerae* has provided findings that are of great pharmacological prominence. The genome level examining procedure including comparison of pathogen and host cellular machinery resulted in identification of a vast set of pathogen specific proteins. Consequent thorough search determined a smaller subset of the functional genome that was essential for pathogen survival. Moreover, it was further annotated to give pathway information. Additionally functional parameters of druggability potential and localization furthered the confidence level in our choice of the most effective drug target. This led to the selection of vibE, as it has a decisive role in formation of vibriobactin, a siderophore from *Vibrio cholerae* necessary for the survival of this bacteria in iron limiting conditions. Efficacious application of the comparative homology modeling generated a high quality model for previously structurally uncharacterized *V. Cholerae* vibE. Rigorous stereochemical verification endorsed the homology model as fit for usage in high level analytical work such as molecular docking and molecular dynamics. Subsequently, it functioned as underpinning for docking studies which revealed 2, 3-dihydroxybenzohydroxamoyl adenylate as strong inhibitor against *V. Cholerae* vibE, exhibiting wide hydrogen bonding. The influence of these interactions was imitated by a GoldScore of 75.7. The strength of polar inhibitor moieties on the level of protein ligand interactions was concluded from these observations, which coincided with the physicochemical character of the binding pocket denoting a preference for charged residues. Molecular dynamic studies including the MD simulations well explained the dynamic behavior of the docked protein. Beside the side chain fluctuations and minor helix to loop movement, stability of inhibitor and target protein complex was observed. Consequently, coupled with information about key functional residues of drug target, the atomic level structural and dynamic insights can fuel a structure based drug design of novel inhibitors with increased drug selectivity, efficacy and specificity against *V. Cholerae.*

References

Albert, M.J., Siddique, A.K., Islam, M.S., Faruque, A.S.G., Ansaruzzaman, M., Faruque, S.M. and Sack, R.B., 1993. Large outbreak of clinical cholera due to Vibrio cholerae non-01 in Bangladesh. *The Lancet*, *341*(8846), p.704.

Allen, M.P. and Tildesley, D.J., 1989. *Computer simulation of liquids*. Oxford university press.

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

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

Arnold, K., Bordoli, L., Kopp, J. and Schwede, T., 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, *22*(2), pp.195-201.

Azam, S.S. and Shamim, A., 2014. An insight into the exploration of druggable genome of Streptococcus gordonii for the identification of novel therapeutic candidates. *Genomics*, *104*(3), pp.203-214.

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

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

Beaber, J.W., Hochhut, B. and Waldor, M.K., 2002. Genomic and functional analyses of SXT, an integrating antibiotic resistance gene transfer element derived from Vibrio cholerae. *Journal of Bacteriology*, *184*(15), pp.4259-4269.

References

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., 2000. The protein data bank. *Nucleic Acids Research*, *28*(1), pp.235-242.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. and Bairoch, A., 2007. Uniprotkb/swiss-prot. In *Plant bioinformatics* (pp. 89-112). Humana Press.

Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X. and Chen, Y.Z., 2003. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Research*, *31*(13), pp.3692-3697.

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

Chemical Computing Group I, 2013. Molecular Operating Environment (MOE).

Colovos, C. and Yeates, T.O., 1993. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Science: A Publication of the Protein Society*, *2*(9), p.1511.

Daniels, N.A., MacKinnon, L., Bishop, R., Altekruse, S., Ray, B., Hammond, R.M., Thompson, S., Wilson, S., Bean, N.H., Griffin, P.M. and Slutsker, L., 2000. Vibrio parahaemolyticus infections in the United States, 1973–1998. *Journal of Infectious Diseases*, *181*(5), pp.1661-1666.

de Sá Morais, L.L.C., Garza, D.R., Loureiro, E.C.B., Nunes, K.N.B., Vellasco, R.S., da Silva, C.P., Nunes, M.R.T., Thompson, C.C., VicvibE, A.C.P. and de Oliveira Santos, E.C., 2012. Complete genome sequence of a sucrose-nonfermenting epidemic strain of Vibrio cholerae O1 from Brazil. *Journal of Bacteriology*, *194*(10), pp.2772-2772.

Duan, Y., Wu, C., Chowdhury, S., Lee, M.C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T. and Caldwell, J., 2003. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry*, *24*(16), pp.1999-2012.

Duffy, J., 1971. The history of Asiatic cholera in the United States. *Bulletin of the New York Academy of Medicine*, *47*(10), p.1152.

References

Dutta, A., Singh, S.K., Ghosh, P., Mukherjee, R., Mitter, S. and Bandyopadhyay, D., 2006. In silico identification of potential therapeutic targets in the human pathogen Helicobacter pylori. *In Silico Biology*, *6*(1), pp.43-47.

Eswar, N., Eramian, D., Webb, B., Shen, M.Y. and Sali, A., 2008. Protein structure modeling with MODELLER. In *Structural Proteomics* (pp. 145-159). Humana Press.

Faruque, S.M., Albert, M.J. and Mekalanos, J.J., 1998. Epidemiology, Genetics, and Ecology of Toxigenic Vibrio cholerae. *Microbiology and Molecular Biology Reviews*, *62*(4), pp.1301-1314.

Feller, S.E., Zhang, Y., Pastor, R.W. and Brooks, B.R., 1995. Constant pressure molecular dynamics simulation: the Langevin piston method. *The Journal of Chemical Physics*, *103*(11), pp.4613-4621.

Franklin, R.B., 2009. In silico studies in ADME/Tox: caveat emptor. *Current Computer-Aided Drug Design*, 5(2), pp.128-138.

Garza-Fabre, M., Toscano-Pulido, G. and Rodriguez-Tello, E., 2012, July. Locality-based multiobjectivization for the HP model of protein structure prediction. In *Proceedings of the 14th annual conference on Genetic and evolutionary computation* (pp. 473-480). ACM.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D. and Bairoch, A., 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, *31*(13), pp.3784-3788.

Gopal, S., Otta, S.K., Kumar, S., Karunasagar, I., Nishibuchi, M. and Karunasagar, I., 2005. The occurrence of Vibrio species in tropical shrimp culture environments; implications for food safety. *International Journal of Food Microbiology*, *102*(2), pp.151-159.

Hayashi, F., Harada, K., Mitsuhashi, S. and Inoue, M., 1982. Conjugation of Drug-Resistance Plasmids from Vibrio anguillarum to Vibrio parahaemolyticus. *Microbiology and Immunology*, *26*(6), pp.479-485.

Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Clayton, R.A., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Umayam, L. and Gill, S.R., 2000. DNA sequence of

both chromosomes of the cholera pathogen Vibrio cholerae. *Nature*, *406*(6795), pp.477-483.

Hughes, J.P., Rees, S., Kalindjian, S.B. and Philpott, K.L., 2011. Principles of early drug discovery. *British Journal of Pharmacology*, *162*(6), pp.1239-1249.

Humphrey, W., Dalke, A. and Schulten, K., 1996. VMD: visual molecular dynamics. *Journal of Molecular Graphics*, *14*(1), pp.33-38.

Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R., 1997. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, *267*(3), pp.727-748.

Kanehisa, M. and Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), pp.27-30.

Knowles, J. and Gromo, G., 2003. Target selection in drug discovery. *Nature Reviews Drug Discovery*, *2*(1), pp.63-69.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V. and Djoumbou, Y., 2011. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research*, *39*(suppl 1), pp.D1035-D1041.

Kovach, M.E., Shaffer, M.D. and Peterson, K.M., 1996. A putative integrase gene defines the distal end of a large cluster of ToxR-regulated colonization genes in Vibrio cholerae. *Microbiology*, *142*(8), pp.2165-2174.

Kumar, V., Chandra, S. and Imran Siddiqi, M., 2014. Recent advances in the development of antiviral agents using computer-aided structure based approaches. *Current Pharmaceutical Design*, *20*(21), pp.3488-3499.

Laskowski, R.A., 2001. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Research*, *29*(1), pp.221-222.

Laskowski, R.A., Moss, D.S. and Thornton, J.M., 1993. Main-chain bond lengths and bond angles in protein structures. *Journal of Molecular Biology*, *231*(4), pp.1049-1067.

References

Lee, S.H., Hava, D.L., Waldor, M.K. and Camilli, A., 1999. Regulation and temporal expression patterns of Vibrio cholerae virulence genes during infection. *Cell*, *99*(6), pp.625-634.

Li, W., Jaroszewski, L. and Godzik, A., 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, *17*(3), pp.282-283.

Li, Z., Wan, H., Shi, Y. and Ouyang, P., 2004. Personal experience with four kinds of chemical structure drawing software: review on ChemDraw, ChemWindow, ISIS/Draw, and ChemSketch. *Journal of Chemical Information and Computer Sciences*, *44*(5), pp.1886-1890.

Lin, W., Fullner, K.J., Clayton, R., Sexton, J.A., Rogers, M.B., Calia, K.E., Calderwood, S.B., Fraser, C. and Mekalanos, J.J., 1999. Identification of a Vibrio cholerae RTX toxin gene cluster that is tightly linked to the cholera toxin prophage. *Proceedings of the National Academy of Sciences*, *96*(3), pp.1071-1076.

Maiorov, V.N. and Crippen, G.M., 1994. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *Journal of Molecular Biology*, *235*(2), pp.625-634.

Martí-Renom, M.A., Stuart, A.C., Fiser, A., Sánchez, R., Melo, F. and Šali, A., 2000. Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, *29*(1), pp.291-325.

May, J.J., Kessler, N., Marahiel, M.A. and Stubbs, M.T., 2002. Crystal structure of DhbE, an archetype for aryl acid activating domains of modular nonribosomal peptide synthetases. *Proceedings of the National Academy of Sciences*, *99*(19), pp.12120-12125.

McCall, M., 2010. *Classical Mechanics: From Newton to Einstein: A Modern Introduction*. John Wiley & Sons.

Mekalanos, J.J., Rubin, E.J. and Waldor, M.K., 1997. Cholera: molecular basis for emergence and pathogenesis. *FEMS Immunology & Medical Microbiology*, *18*(4), pp.241-248.

References

Meng, X.Y., Zhang, H.X., Mezei, M. and Cui, M., 2011. Molecular docking: a powerful approach for structure-based drug discovery. *Current Computer-Aided Drug Design*, *7*(2), p.146-157.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. and Kanehisa, M., 2007. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, *35*(suppl 2), pp.W182-W185.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G. and Thornton, J.M., 1992. Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Bioinformatics*, *12*(4), pp.345-364.

Nancy, Y.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J. and Brinkman, F.S., 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, *26*(13), pp.1608-1615.

Notredame, C., Higgins, D.G. and Heringa, J., 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, *302*(1), pp.205-217.

Parvege, M.M., Rahman, M. and Hossain, M.S., 2014. Genome-wide Analysis of Mycoplasma hominis for the Identification of Putative Therapeutic Targets. *Drug target insights*, *8*, p.51.

Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.S., Cheatham, T.E., DeBolt, S., Ferguson, D., Seibel, G. and Kollman, P., 1995. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, *91*(1), pp.1-41.

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

References

Piddock, L.J., 2006. Clinically relevant chromosomally encoded multidrug resistance efflux pumps in bacteria. *Clinical Microbiology Reviews*, *19*(2), pp.382-402.

Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M.S., Davis, F.P., Stuart, A.C., Mirkovic, N., Rossi, A., Marti-Renom, M.A., Fiser, A. and Webb, B., 2004. MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Research*, *32*(suppl 1), pp.D217-D222.

Ramakrishnan, C. and Ramachandran, G.N., 1965. Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophysical Journal*, *5*(6), pp.909-933.

Reddy, R.N., Mutyala, R., Aparoy, P., Reddanna, P. and Reddy, M.R., 2007. Computer aided drug design approaches to develop cyclooxygenase based novel anti-inflammatory and anti-cancer drugs. *Current Pharmaceutical Design*, *13*(34), pp.3505-3517.

Rusnak, F., Faraci, W.S. and Walsh, C.T., 1989. Subcloning, expression, and purification of the vibErobactin biosynthetic enzyme 2, 3-dihydroxybenzoate-AMP ligase: demonstration of enzyme-bound (2, 3-dihydroxybenzoyl) adenylate product. *Biochemistry*, *28*(17), pp.6827-6835.

Russ, A.P. and Lampel, S., 2005. The druggable genome: an update. *Drug discovery today*, *10*(23), pp.1607-1610.

Ryckaert, J.P., Ciccotti, G. and Berendsen, H.J., 1977. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, *23*(3), pp.327-341.

Sarangi, A.N., Aggarwal, R., Rahman, Q. and Trivedi, N., 2009. Subtractive genomics approach for in silico identification and characterization of novel drug targets in Neisseria Meningitides Serogroup B. *Journal of Computer Science & Systems Biology*, *2*(5), pp.255-258.

Schomburg, I., Chang, A. and Schomburg, D., 2002. BRENDA, enzyme data and metabolic information. *Nucleic Acids Research*, *30*(1), pp.47-49.

References

Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C., 2003. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research*, *31*(13), pp.3381-3385.

Shimada, T., Arakawa, E., Itoh, K., Okitsu, T., Matsushima, A., Asai, Y., Yamai, S., Nakazato, T., Nair, G.B., Albert, M.J. and Takeda, Y., 1994. Extended serotyping scheme for Vibrio cholerae. *Current Microbiology*, *28*(3), pp.175-178.

Slamti, L., Livny, J. and Waldor, M.K., 2007. Global gene expression and phenotypic analysis of a Vibrio cholerae rpoH deletion mutant. *Journal of Bacteriology*, *189*(2), pp.351-362.

Trott, O. and Olson, A.J., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, *31*(2), pp.455-461.

Van Heyningen, W.E. and Seal, J.R., 1983. *Cholera. The American Scientific Experience 1947-1980*. Bowker Publishing Co.

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

Visualizer, D.S., 2012. Release 3.5. *Accelrys Inc, San Diego, CA, USA.*

Waldor, M.K. and Mekalanos, J.J., 1996. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science*, *272*(5270), pp.1910-1914.

Wallace, A.C., Laskowski, R.A. and Thornton, J.M., 1995. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Engineering*, *8*(2), pp.127-134.

Weiner, P.K. and Kollman, P.A., 1981. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *Journal of Computational Chemistry*, *2*(3), pp.287-303.

Wereszczynski, J. and McCammon, J.A., 2012. Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. *Quarterly Reviews of Biophysics*, *45*(1), pp.1-25.

References

Wiederstein, M. and Sippl, M.J., 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research*, *35*(suppl 2), pp.W407-W410.

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

Yamaichi, Y., Iida, T., Park, K.S., Yamamoto, K. and Honda, T., 1999. Physical and genetic map of the genome of Vibrio parahaemolyticus: presence of two chromosomes in Vibrio species. *Molecular Microbiology*, *31*(5), pp.1513-1521.

Zhang, R. and Lin, Y., 2009. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Research*, *37*(suppl 1), pp.D455-D458.

Zhang, Y., 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, *9*(1), p.40.

Appendices

Appendix 1: Pathway annotation for *V. Cholerae* LMA3984-4 proteins through KAAS.

Gene Name	Protein Name	EC Number	Pathway Name
PTS-EI.PTSI	phosphotransferase system, enzyme I, PtsI	EC:2.7.3.9	Phosphotransferase system
gspE	general secretion pathway protein E	-	Bacterial secretion system
torD	TorA specific chaperone	-	Two-component system
glnG	two-component system, NtrC family, nitrogen regulation response regulator GlnG	-	Two-component system
nifa	Nif-specific regulatory protein	-	Two-component system

***Remaining Appendix 1 for three strains is provided in soft copy.**

Appendix 2: Drug Bank assessment of *V. Cholerae* LMA3984-4 proteins

Protein Name	Target Name	Bit Score	Drug Bank ID
7-Keto-8-Aminopelar-gonic Acid	Adenosylmethionine-8-amino-7-oxononano-ate aminotransferase	489.189	DB02274
Dihydroorotic Acid	Dihydroorotase	380.948	DB02129
Isocitric Acid	Isocitrate dehydrogen-ase [NADP]	1070	DB01727
2,3-Dihydroxy-Ben-zoic Acid	2,3-dihydroxybenzo-ate-AMP ligase	584.719	DB01672
Indole-3-Propanol Phosphate	Tryptophan synthase alpha chain	673.315	DB03171

***Remaining Appendix 2 for three strains is provided in soft copy.**

Thesis

ORIGINALITY REPORT

6%	3%	5%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mswg.org.my Internet Source	<1%
2	Kaur, Navkiran, Mansimran Khokhar, Vaibhav Jain, P. V. Bharatam, Rajat Sandhir, and Rupinder Tewari. "Identification of Druggable Targets for Acinetobacter baumannii Via Subtractive Genomics and Plausible Inhibitors for MurA and MurB", Applied Biochemistry and Biotechnology, 2013. Publication	<1%
3	Wang, L.. "Ethaselen: a potent mammalian thioredoxin reductase 1 inhibitor and novel organoselenium anticancer agent", Free Radical Biology and Medicine, 20120301 Publication	<1%
4	Fracchiolla, G.. "Synthesis, biological evaluation, and molecular modeling investigation of chiral 2-(4-chloro-phenoxy)-3-phenyl-propanoic acid derivatives with PPAR@a and PPAR@c agonist activity",	<1%

Bioorganic & Medicinal Chemistry, 20081101
Publication

---

5 Azam, S. Sikander, and A. Hammad Mirza. "Role of thumb index fold in Wnt-4 protein and its dynamics through a molecular dynamics simulation study", Journal of Molecular Liquids, 2014. <1%
Publication

---

6 Dominguez, C.. "Structure determination and dynamics of protein-RNA complexes by NMR spectroscopy", Progress in Nuclear Magnetic Resonance Spectroscopy, 201102 <1%
Publication

---

7 Dukhovskoy, Dmitry S., Jonathan Ubnoske, Edward Blanchard-Wrigglesworth, Hannah R. Hiester, and Andrey Proshutinsky. "Skill metrics for evaluation and comparison of sea ice models", Journal of Geophysical Research Oceans, 2015. <1%
Publication

---

8 Azam, S. Sikander, Reaz Uddin, and Abdul Wadood. "Structure and dynamics of alpha-glucosidase through molecular dynamics simulation studies", Journal of Molecular Liquids, 2012. <1%
Publication

---

9 www.researchgate.net

Internet Source

<1%

---

10 Ebrahimi, Malihe, and Taghi Khayamian. "Interactions of G-quadruplex DNA binding site with berberine derivatives and construct a structure-based QSAR using docking descriptors", Medicinal Chemistry Research, 2014.
Publication

<1%

---

11 Azam, Syed Sikander, Asma Abro, Farya Tanvir, and Nousheen Parvaiz. "Identification of unique binding site and molecular docking studies for structurally diverse Bcl-xL inhibitors", Medicinal Chemistry Research, 2014.
Publication

<1%

---

12 Azam, Syed Sikander, Asma Abro, Saad Raza, and Ayman Saroosh. "Structure and dynamics studies of sterol 24-C-methyltransferase with mechanism based inactivators for the disruption of ergosterol biosynthesis", Molecular Biology Reports, 2014.
Publication

<1%

---

13 www.scribd.com
Internet Source

<1%

---

14 www.coursehero.com
Internet Source

<1%

15	Trucksis, Michele Michalski, Jane Deng, . "The Vibrio cholerae genome contains two unique circular chromosomes.", Proceedings of the National Academy of S, Nov 24 1998 Issue Publication	<1%
16	nar.oxfordjournals.org Internet Source	<1%
17	www.biomodelit.com Internet Source	<1%
18	Politi, A.. "Development of accurate binding affinity predictions of novel renin inhibitors through molecular docking studies", Journal of Molecular Graphics and Modelling, 201011 Publication	<1%
19	Abu-Bakar, A.. "Metabolism of bilirubin by human cytochrome P450 2A6", Toxicology and Applied Pharmacology, 20120515 Publication	<1%
20	www.stat.duke.edu Internet Source	<1%
21	www.msg.ucsf.edu Internet Source	<1%
22	Azam, Syed Sikander, Sumra Wajid Abbasi, Amina Saleem Akhtar, and Mah-laka Mirza. "Comparative modeling and molecular docking	<1%

studies of d-Alanine:d-alanine ligase: a target of antibacterial drugs", Medicinal Chemistry Research, 2014.
Publication

---

23 www.pcst.org.pk
Internet Source
<1%

---

24 www.ptfarm.pl
Internet Source
<1%

---

25 Azam, Syed Sikander, and Saad Raza. "Structure modeling and hybrid virtual screening study of Alzheimer's associated protease kallikrein 8 for the identification of novel inhibitors", Medicinal Chemistry Research, 2014.
Publication
<1%

---

26 Gogoi, Prerana, Monika Chandravanshi, Suraj Kumar Mandal, Ambuj Srivastava, and Shankar Prasad Kanaujia. "Heterogeneous behavior of metalloproteins toward metal ion binding and selectivity: insights from molecular dynamics studies", Journal of Biomolecular Structure and Dynamics, 2015.
Publication
<1%

---

27 Azam, Syed Sikander, and Amen Shamim. "An insight into the exploration of druggable genome of Streptococcus gordonii for the identification of novel therapeutic candidates",
<1%

Genomics, 2014.
Publication

28 Li, A.. "Modal reduction of mathematical models of biological molecules", Journal of Computational Physics, 20060101 <1%
Publication

29 Toyama, Daniela de Oliveira Gaeta, Henri. "An evaluation of 3-rhamnosylquercetin, a glycosylated form of quercetin, against the myotoxic and ed", BioMed Research International, Annual 2014 Issue <1%
Publication

30 Doss, C. George Priya Chakraborty, Chira. "Integrating in silico prediction methods, molecular docking, and molecular dynamics simulation to pr", BioMed Research International, Annual 2014 Issue <1%
Publication

EXCLUDE QUOTES	ON	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	ON		