

**COMPUTATIONAL STUDY OF ANTIBIOTIC RESISTANCE
MECHANISMS IN DIARRHEAL PATHOGENS AND
IDENTIFICATION OF POTENTIAL DRUG TARGETS AND
VACCINE CANDIDATES**

**A THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY**

**IN
BIOINFORMATICS**

**BY
KUSUM
Enrollment No. 136503**



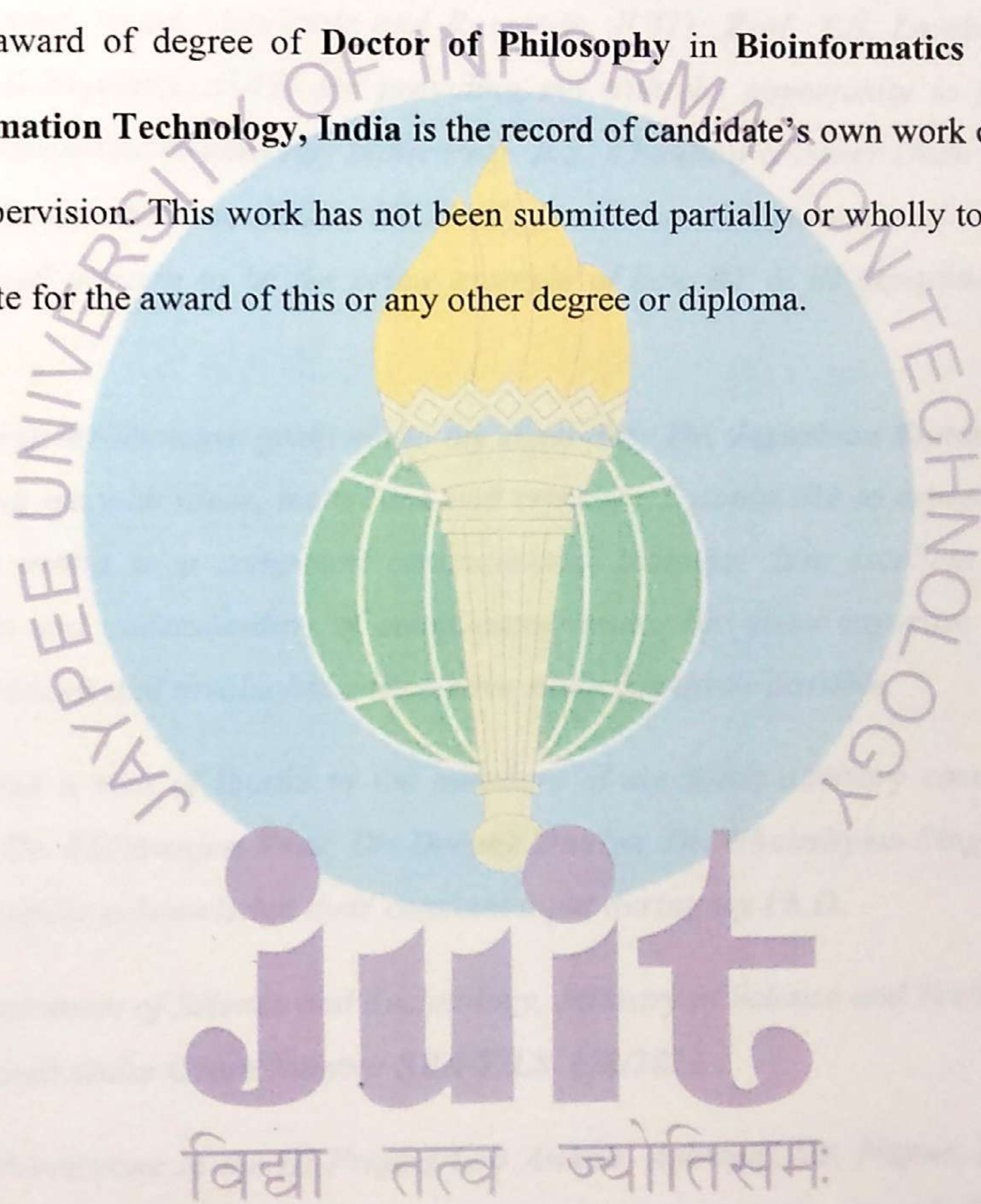
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT

JUNE, 2017

CERTIFICATE

This is to certify that the thesis entitled, “Computational Study of Antibiotic Resistance Mechanisms in Diarrheal Pathogens and Identification of Potential Drug Targets and Vaccine Candidates” which is being submitted by **Kusum** (Enrollment No. 136503) in fulfillment for the award of degree of **Doctor of Philosophy** in **Bioinformatics** at **Jaypee University of Information Technology, India** is the record of candidate’s own work carried out by her under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.



J. Ramana

Dr. Jayashree Ramana

Assistant Professor (Senior Grade),

Email: jayashree.ramana13@gmail.com

Date: 7/6/2017

DECLARATION

I certify that:

- The work contained in this thesis is original and has been done by me under the guidance of my supervisor.
- The work has not been submitted to any other organisation for any degree or diploma.
- Wherever, I have used materials (data, analysis, figures or text), I have given due credit by citing them in the text of the thesis.



Kusum
Kusum

(Enrollment No. 136503)

Department of Biotechnology and Bioinformatics

Jaypee University of Information Technology, Waknaghat, India

Date: 07/06/2017

TABLE OF CONTENTS

DECLARATION	III
CERTIFICATE	IV
ACKNOWLEDGMENT	VI
LIST OF FIGURES	XIII
LIST OF TABLES	XX
ABBREVIATIONS	XXIII
ABSTRACT	XXVII

CHAPTER 1

INTRODUCTION	1– 25
1.1 INTRODUCTION.....	2
1.2 DIARRHEAL ETIOLOGY.....	6
1.2.1 Bacteria.....	6
1.2.1.1 <i>Escherichia coli</i>	6
1.2.1.2 <i>Vibrio Species</i>	7
1.2.1.3 <i>Shigella Species</i>	7
1.2.1.4 <i>Campylobacter Species</i>	8
1.2.1.5 <i>Salmonella Species</i>	8
1.2.1.6 <i>Clostridium Species</i>	8
1.2.1.7 <i>Yersinia enterocolitica</i>	9
1.2.1.8 <i>Bacillus cereus</i>	9
1.2.2. Viruses	9

1.2.3 Parasites.....	10
1.3 DIAGNOSIS	11
1.4 TREATMENT	11
1.5 ANTIMICROBIAL RESISTANCE (AMR).....	13
1.5.1 Antibiotic resistance mechanisms	14
1.5.2 Status quo of antibiotic resistance in diarrheal pathogens.....	16
1.6 KNOWLEDGE GAPS	17
1.9 OBJECTIVES.....	18
REFERENCES.....	21

CHAPTER 2

<i>DBDiaSNP: An Open-Source Knowledgebase of Genetic Polymorphisms and Resistance Genes related to Diarrheal Pathogens</i>	26-37
2.1 INTRODUCTION.....	27
2.2 METHODS.....	28
2.2.1 Database system implementation	28
2.2.1 SNP and resistance data search	29
2.3 RESULTS AND DISCUSSION.....	29
2.3.1 Development of SNP-based diagnostics.....	29
2.3.2 SNPs monitoring in different geographical regions.....	30
2.3.3 Drug discovery.....	32
2.4 FUTURE DIRECTIONS	34
CONCLUSION.....	34
REFERENCES.....	35

CHAPTER 3

<i>Molecular Dynamics Simulations and Residue Interaction Network Analysis in Enterotoxigenic E. coli and C. jejuni gyrA mutants: Repercussions on Quinolone Affinity</i>	38-74
--	--------------

3.1 INTRODUCTION.....	39
3.2 METHODS.....	41
3.2.1 Ligand preparation	42
3.2.2 Homology modeling	42
3.2.3 Molecular docking	43
3.2.4 MD simulations	44
3.2.5 MM-PBSA calculations	45
3.2.6 Principal component analysis (PCA).....	46
3.2.7 Residue Interaction Networks (RINs) analysis	46
3.3 RESULTS AND DISCUSSION.....	47
ETEC	
3.3.1 Model building and structure validation	47
3.3.2 Molecular docking and binding affinity calculations	47
3.3.3 Structural stability characterization	51
3.3.4 Secondary structure content	52
3.3.5 MM-PB(GB)/SA binding free energy calculations	54
3.3.6 Per-residue decomposition of binding energy.....	56
3.3.7 PCA	56
3.3.8 RIN analysis	58
<i>C. jejuni</i>	
3.3.9 Model building and structure validation	58
3.3.10 Molecular docking and binding affinity calculations.....	59
3.3.11 Structural stability characterization.....	62
3.3.12 Secondary structure content.....	63
3.3.13 MM-PB(GB)/SA binding free energy calculations	65
3.3.14 Per-residue decomposition of binding energy.....	65
3.3.15 PCA	67
3.3.16 RIN analysis	69
CONCLUSION.....	71
REFERENCES.....	71

CHAPTER 4

Surface Proteome Mining for Identification of Potential Vaccine Candidates against Campylobacter jejuni: an in silico approach 75-93

4.1 INTRODUCTION.....	76
4.2 METHODS.....	77
4.2.1 Retrieving non-homologous proteins from pathogen whole proteome.....	77
4.2.2 Antigenicity and transmembrane prediction.....	77
4.2.3 T-cell epitope prediction	78
4.2.4 Epitope conservancy and HLA-distribution analysis	79
4.2.5 Molecular docking studies of HLA-epitope.....	79
4.2.5.1 Designing epitope 3D structure	79
4.2.5.2 Docking.....	80
4.2.6 B-cell epitope identification.....	80
4.3 RESULTS.....	80
4.3.1 Retrieving non-homologous proteins from pathogen whole proteome	80
4.3.2 Antigenicity and transmembrane prediction	81
4.3.3 T-cell epitope prediction	81
4.3.4 Epitope conservancy and HLA-distribution analysis	81
4.3.5 Molecular docking studies of HLA-epitope.....	82
4.3.6 B-cell epitope identification.....	84
4.4 DISCUSSION.....	85
CONCLUSION.....	89
REFERENCES.....	90

CHAPTER 5

Identification of Epitope-Based Peptide Vaccine Candidates against Enterotoxigenic Escherichia coli: A Comparative Genomics and Immunoinformatics Approach 94-117

5.1 INTRODUCTION.....	95
5.2 METHODS.....	96
5.2.1 Comparative genomics analysis	96

5.2.2 Antigenicity and transmembrane prediction.....	98
5.2.3 T-cell epitope prediction	98
5.2.4 HLA-distribution analysis	98
5.2.5 Molecular docking studies of HLA-epitope.....	79
5.2.5.1 Designing epitope 3D structure	99
5.2.5.2 Docking.....	99
5.2.6 B-cell epitope identification.....	99
5.3 RESULTS.....	100
5.3.1 Genomic Features.....	100
5.3.2 Comparative genomics analysis	101
5.3.3 Antigenicity and transmembrane prediction	102
5.3.4 T-cell epitope prediction	102
5.3.5 HLA-distribution analysis	103
5.3.6 Molecular docking studies of HLA-epitope.....	104
5.3.7 B-cell epitope identification.....	105
5.4 DISCUSSION.....	107
CONCLUSION.....	114
REFERENCES.....	114

CHAPTER 6

Novel Drug Targets for Food-Borne Pathogen Campylobacter jejuni: An Integrated Subtractive Genomics and Comparative Metabolic Pathway Study 118-138

6.1 INTRODUCTION.....	119
6.2 METHODS.....	120
6.2.1 Host and pathogen comparative metabolic pathway analysis.....	121
6.2.2 Subtractive genomics and identification of essential non-homologous pathogen proteins.....	121
6.2.3 Prioritization of essential non-homologous proteins for therapeutic targets.....	121
6.2.4 Druggability of essential non-homologous proteins.....	122
6.3 RESULTS AND DISCUSSION.....	122

6.3.1 Identification of unique metabolic pathways.....	122
6.3.2 Identification of essential non-homologous pathogen proteins.....	123
6.3.3 Prioritization of essential non-homologous proteins for therapeutic drug targets.....	124
6.3.4 Druggability of potential drug targets.....	125
CONCLUSION.....	135
REFERENCES.....	135

CHAPTER 7

Tapping into Salmonella typhimurium LT2 genome in a quest to explore its therapeutic arsenal: a metabolic network modeling approach 139-163

7.1 INTRODUCTION.....	140
7.2 METHODS.....	141
7.2.1 Gene knockout analysis	142
7.2.2 Identification of essential non-homologous pathogen proteins	142
7.2.3 Prioritization of drug targets	143
7.2.4 Druggability and assayability	143
7.2.5 Protein-Protein Interaction (PPI) network analysis.....	144
7.3 RESULTS AND DISCUSSION.....	145
7.3.1 Gene knockout analysis	145
7.3.2 Identification of essential non-homologous pathogen proteins.....	146
7.3.3 Prioritization of drug targets	146
CONCLUSION.....	159
REFERENCES.....	160
CONCLUSION AND FUTURE PERSPECTIVES.....	164-167
ANNEXURE.....	168-171
PUBLICATIONS AND CONFERENCES.....	172-174

LIST OF FIGURES

Figure 1.1: Rates of child mortality due to diarrhea.

Figure 1.2: Regional burden of diarrhea per year in children aged 0–4 years, by WHO region.

Figure 1.3: Causes of deaths in under-five children in India.

Figure 1.4: Pathways to diarrheal spread through contaminated water, food, or objects.

Figure 1.5: Distribution pattern of causes of diarrheal death in a) children below age 5 and b) all age groups globally from 1999-2013.

Figure 1.6: Timeline of the discovery of antibiotic classes showing the dearth of progress made in the last few decades.

Figure 1.7: The various mechanisms of antibiotic resistance.

Figure 1.8: Percentage of isolates of different diarrheal pathogens resistant to antimicrobial agents, 2013.

Figure 2.1: DBDiaSNP architecture overview. a) Standard three-tier architecture of DBDiaSNP; b) Organizational structure of the DBDiaSNP. The data were compiled from existing literature and online databases Uniprot, ClinVar, dbSNP, and OMIM, and enriched with annotation information such as PDB ID, Gene ID, Pubmed ID, Biological function, and CCDS ID from online resources. The database includes three modules: 1) Polymorphism data further categorized into Host and Pathogen Polymorphism; 2) Resistance Gene data further categorized into Plasmid mediated, Integrons, and Gene Cassettes, and Transposon Mediated; 3) Research articles relevant to each category.

Figure 2.2: Typical output of a query (SNP). (a) Description of different annotation terms; (b) Distribution of mutations associated with antibiotic resistance in *C. difficile*; (c) Mutation in IL8 associated with susceptibility to EAEC-associated traveler's diarrhea in *H. sapiens*.

Figure 2.3: Typical output of a query (Resistance Gene). This view shows the distribution of resistance genes distributed in *Campylobacter* spp.

Figure 3.1: Cartoon representation of ETEC gyrA showing its distinctive domains, catalytic site and mutations associated with quinolone resistance reported in diarrheal patients. QRDR mutations and the catalytic site are in stick representation and shown in yellow and cyan color respectively. The CAP domain is colored violet purple, the tower blue, the dimerization domain and connecting α -helices colored orange. Zoomed in is a view of CAP domain showing HTH motif with three helices and two beta-sheets with corresponding residue ranges as α 1 (43–56), α 2 (66–77), α 3 (81–92), β 2 (101–106), β 3 (124–128).

Figure 3.2: Superimposition of ETEC gyrA and the template (PDB ID: 1AB4) produced by structural modeling with color coding scheme for helix, sheet and loop as red, yellow, green in template structure and cyan, magenta, salmon in the modeled structure.

Figure 3.3: Highest scoring pose of ciprofloxacin at active site of ETEC WT gyrA. Interacting residues and atoms are labeled. Receptor is shown as cartoons. Ligand is shown in stick mode, gyrA interacting residues are shown in ball and stick mode. Binding interactions are shown as dashed lines indicating salt bridge and hydrogen bonding interactions.

Figure 3.4: Hydrogen bonds involved in binding of ciprofloxacin with a) wild type b) mutated S83L and c) S83L/D87N QRDR of gyrA.

Figure 3.5: Time evolution RMSD of gyrA backbone throughout the 10 ns MD simulation time for WT (Blue), S83L mutant (Yellow), S83L/D87N mutant (Red).

Figure 3.6: Time evolution RMSF of gyrA backbone atoms throughout the 10 ns MD simulation time for WT (Blue), S83L mutant (Yellow), S83L/D87N mutant (Red).

Figure 3.7: Average secondary-structure contents of protein backbone residues in a) WT, b) S83L mutant, and c) and S83L/D87N double mutant.

Figure 3.8: Observed conformational changes of catalytic residue Tyr 122 and the mutated residues Ser 83 and Asp 87 during the course of MD simulation in WT (Column 1), S83L mutant (Column 2), S83L/D87N mutant (Column 3) at 0 ns (start of production phase), 5 ns, and 10 ns (end of production phase).

Figure 3.9: Time evolution of rotation of dihedral bond angles a) dihedral angle phi b) dihedral angle psi throughout MD simulation course in WT, S83L mutant and S83L/D87N mutant. Relatively stable dihedral angle of catalytic residue Tyr 122 across all structures throughout simulations suggests no major conformational change.

Figure 3.10: Polar and non-polar contribution of hot-spot residues from MM-PBSA based energy decomposition analysis of total binding energy (kcal/mol) during last 5 ns of MD simulation in a) WT, b) S83L mutant, and c) S83L/D87N mutant.

Figure 3.11: Porcupine plot of the first eigenvector obtained by PCA from simulation ensembles of (a) WT, (b) S83L mutant, and S83L/D87N mutant from the last 5 ns of MD simulation. Protein backbone represented as ribbon. Dominant motions of residues illustrated as arrows. Arrow represents direction of the eigenvector and the size of each arrow shows the magnitude of the corresponding eigenvalue.










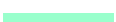

Figure 3.12: Residue interaction networks (RINs) comparison between a) WT, b) S83L mutant, and c) S83L/D87N mutant docked with ciprofloxacin. Zoomed view highlights changes in network in vicinity of catalytic residue Tyr 122 and mutated QRDR residues Ser 83 and Asp 87. Nodes represented as: Helix  Loop  Sheet  Default  and Edges represented as: Hydrogen Bond  Peptide Bond  Salt Bridge  Ionic bond  Pi Stacking  Pi-Cation  Closest Atoms Interaction 

Figure 3.13 Structural superimposition of *C. jejuni* gyrA on the template (PDB ID: 1AB4) produced by structural modeling with modeled structure colored cyan and template colored sea green.

Figure 3.14: Highest scoring pose of ciprofloxacin at active site of ETEC WT gyrA. Interacting residues and atoms are labeled. Receptor is shown as cartoons. Ligand is shown in stick mode, gyrA interacting residues are shown in ball and stick mode. Binding interactions are shown as dashed lines indicating salt bridge and hydrogen bonding interactions.

Figure 3.15: Time evolution RMSD of gyrA backbone throughout the 10 ns MD simulation time for WT (Blue), T86I mutant (Yellow), T86I/P104S mutant (Red).

Figure 3.16: Time evolution RMSF of gyrA backbone atoms throughout the 10 ns MD simulation time for WT (Blue), T86I mutant (Yellow), T86I/P104S mutant (Red).





Figure 3.17: Average secondary-structure contents of protein backbone residues in a) WT, b) T86I mutant, and c) and T86I/P104S double mutant.




Figure 3.18: Observed conformational changes of the protein backbone and mutated residue positions 86 and 104 during the course of MD simulation in WT (Row 1), T86I mutant (Row 2), T86I/P104S mutant (Row 3) at 0 ns (start of production phase), 5 ns, and 10 ns (end of production phase).

Figure 3.19: Polar and non-polar contribution of hot-spot residues from MM-PBSA based energy decomposition analysis of total binding energy (kcal/mol) in a) WT, b) T86I mutant, and c) and T86I/P104S double mutant during last 5 ns of MD simulation.

Figure 3.20: Porcupine plot of the first eigenvector obtained by PCA from simulation ensembles of a) WT, b) T86I mutant, and c) and T86I/P104S mutant from the last 5 ns of MD simulation. Protein backbone represented as ribbon. Dominant motions of residues illustrated as arrows. Arrow represents direction of the eigenvector and the size of each arrow shows the magnitude of the corresponding eigen value.

Figure 3.21: Residue interaction networks (RINs) comparison between a) WT, b) T86I mutant, and c) and T86I/P104S mutant docked with ciprofloxacin. Zoomed view highlights changes in network in the vicinity of mutated residue.

Nodes represented as: Helix  Loop  Sheet  Default  and

Edges represented as: Hydrogen Bond  Peptide Bond  Salt Bridge 





Ionic bond  Pi Stacking  Pi-Cation  Closest Atoms Interaction 

Figure 4.1: Schematic representation of the protocols used for epitope identification.

Figure 4.2: Docked complexes of HLA-A*11:01 against predicted epitopes generated by Hex docking program. a) Epitope NTDQAQGTV b) Epitope YIQDNFNFY.

Figure 4.3: Interactions involved in HLA-A*11:01 binding with a) Epitope NTDQAQGTV b) Epitope YIQDNFNFY.

Figure 4.4: Step-wise reduction in the total no of proteins in search for the identification of vaccine candidates against *C. jejuni*.

Figure 5.1: Schematic representation of the steps involved in target protein prioritization and antigenic epitope mapping in ETEC E24377A strain using an integrated comparative genomics and immunoinformatics approach.

Figure 5.2: Circular genome of ETEC E24377A.

(i) ORFs plus strand: The outer most ring in black color shows ORFs plus strand. In the same ring orange regions indicate the ORFs negative strand (ii) Subcellular localization: In the adjacent ring, green, yellow, purple, light blue, orange, red, and gray color represents cytoplasm, cytoplasmic membrane, outer membrane, extracellular region, cell wall, periplasmic region and unknown subcellular localization respectively (iii) The third red ring represents GC percentage. Peaks represent an increase in GC percentage (iv) fourth ring in yellow color shows GC skew values for the whole genome. (v) The innermost purple colored ring shows AT skew values calculated for the whole genome. The figure is generated by Circular Visualization for microbial genomes (CiVi).

Figure 5.3: Representation of 3 prioritized target proteins: The proteins characterized as i) shared between pathogenic strains and excluded from commensal strain ii) Non-homologous to humans iii) Antigenic iv) Membrane localize, are shown using different colors in the Venn diagram. Proteins satisfying a particular parameter are shown in the corresponding category of the Venn diagram. Three proteins were prioritized for vaccine candidate identification. The image has been generated by Jvenn.

Figure 5.4: A) Docked complex of HLA-A*11:01 with epitope LICFFTLSY visualized in Pymol A') and corresponding interactions involved in binding visualized in LigPlot.

Figure 5.5: B) Docked complex of HLA-A*11:01 with epitope PIVNLFLLY visualized in Pymol B') and corresponding interactions involved in binding visualized in LigPlot.

Figure 5.6: C) Docked complex of HLA-A*11:01 with epitope PLNPLILLY visualized in Pymol C') and corresponding interactions involved in binding visualized in LigPlot.

Figure 5.7: D) Docked complex of HLA-A*11:01 with epitope SVFIFLFIY visualized in Pymol D') and corresponding interactions involved in binding visualized in LigPlot.

Figure 5.8: E) Docked complex of HLA-A*11:01 with epitope SVSVFIFLF visualized in Pymol E') and corresponding interactions involved in binding visualized in LigPlot.

Figure 6.1 Schematic workflow of comparative genomics approach to identify drug targets in pathogenic *C. jejuni* strain NCTC11168.

Figure 6.2: Distribution of essential non-homologous proteins of *C. jejuni* in different pathways before and after drug bank search for druggability potential evaluation.

Figure 6.3 Step-wise reduction in the total no of proteins in the subtractive analysis of *C. jejuni* proteome for drug target identification

Figure 7.1: Circular genome of *S. typhimurium* LT2

(i) ORFs plus strand: The outer most ring shows ORFs plus strand. (ii) The second ring indicates the ORFs negative strand. In two outermost rings green, yellow, purple, light blue, orange, red, and gray color represent cytoplasm, cytoplasmic membrane, outer membrane, extracellular region, cell wall, periplasmic region and unknown subcellular localization respectively (iii) The third red ring in orange represents GC percentage. Peaks represent an increase in GC percentage (iv) fourth ring in blue color shows GC skew values for the whole genome. (v) The innermost red colored ring shows AT skew values calculated for the whole genome. The figure is generated by Circular Visualization for microbial genomes (CiVi).

Figure 7.2: Schematic representation of the steps involved in target protein prioritization from genome scale metabolic network of *S. typhimurium* LT2.

Figure 7.3: Distribution of the essential target proteins retrieved using FBA and MOMA approach in different cellular compartments.

Figure 7.4: Representation of 47 prioritized target proteins: The proteins characterized as i) proteins with a role in metabolic network ii) essential proteins using combined predictions of FBA and MOMA approach iii) Non-homologous to humans iv) experimentally predicted as essential from prediction of DEG v) druggable, are shown using different colors in the Venn

diagram. Proteins satisfying a particular parameter are shown in the corresponding category of the Venn diagram. Three proteins were prioritized as drug targets. The image has been generated by Jvenn.

Figure 7.5: Distribution of essential non-homologous proteins of *S. typhimurium* LT2 in different pathways before and after drug bank search for druggability potential evaluation.

Figure 7.6: Protein–protein interactions among the prioritized proteins. Proteins involved in peptidoglycan synthesis, folate synthesis show extensive linkages among them. Proteins involved in L-rhamnose synthesis are rich in interactions among themselves.

Figure 7.7: 3D structures of the prioritized proteins a) pabA, b) glmM, c) glmU, d) mraY, e) rfbF, and f) rfbH for which no crystal structure information was available based on novel nonlinear context-specific alignment.

Figure 7.8: Final Step catalyzed by rfbD in dTDP-L-rhamnose biosynthesis pathway viz. reduction of dTDP-6-deoxy-L-lyxo-4-hexulose to dTDP-L-rhamnose.

Figure 7.9: A critical node in folate synthesis, Dihydrofolate reductase, links folate synthesis to tetrahydrofolate production (THF) utilization by reducing dihydrofolate to tetrahydrofolate.

LIST OF TABLES

Table 1.1: Classification of diarrhea based on clinical presentations.

Table 1.2: Classification of diarrhea based on intestinal fluid absorbance capacity .

Table 1.3 List of currently available drugs and vaccines against diarrheal pathogens.

Table 2.1: SNPs statistics along with genes: Species wise distribution in pathogens.

Table 2.2: Resistance Gene statistics: Species wise distribution in pathogens.

Table 2.3: SNPs statistics along with genes: Species wise distribution in hosts.

Table 2.4: Breakup of Source of SNPs in database.

Table 3.1: Docking scores and hydrogen bond residues involved in ciprofloxacin binding with WT and MT gyrA.

Table 3.2: Hydrogen bonds involved in gyrA-ciprofloxacin interaction in WT and MT gyrA.

Table 3.3: Contribution of various energy terms in ciprofloxacin binding with WT and MT forms of gyrA .

Table 3.4: MM-PB(GB)/SA based free binding energy profile of ciprofloxacin complexed with the wild and the S83L, S83L/D87N mutant types of gyrA.

Table 3.5: Docking scores and hydrogen bond residues involved in ciprofloxacin binding with WT and mutant forms of gyrA.

Table 3.6: MM-PB(GB)/SA based free binding energy profile of ciprofloxacin complexed with the wild and the S83L, S83L/D87N mutant types of gyrA.

Table 4.1: Most probable predicted epitopes selected on the basis of their NetCTL (MHC binding, proteasomal processing and TAP transport), AllerHunter (Allergic cross-reactivity) and VaxiJen (Antigenicity) Score .

Table 4.2: Predicted potential T cell epitopes, along with their interacting MHC-I and MHC-II alleles with an affinity of <500 nM and corresponding IC50 values (bracketed).

Table 4.3: Population coverage of predicted epitopes based on MHC-I and MHC-II restriction data. **4.3a)** for epitope NTDQAQGTV maximum population coverage by Europe **4.3b)** for epitope YIQDNFNFY maximum population coverage by Europe.

Table 4.4: Four most potential B-cell epitopes by combined predictions of AAP, BCPred, IEDB tools (Emini Surface Accessibility, Karplus and Schulz flexibility, Parker hydrophilicity), filtered based on their AllerHunter and VaxiJen score.

Table 5.1: Most probable predicted epitopes selected on the basis of their NetCTL (MHC binding, proteasomal processing and TAP transport) and VaxiJen (Antigenicity) Score. Epitopes are categorized into High, Medium and Low priority epitopes based on antigenicity scale (VaxiJen Score). Antigenicity Scale: **0.7-1.0:** High Priority **1.0-1.7:** Medium Priority **1.7 and above:** High Priority.

Table 5.2: Predicted potential T cell epitopes, along with their interacting MHC-I and MHC-II alleles with an affinity of <500 nM and corresponding IC50 values (bracketed).

Table 5.3: Population coverage of predicted epitopes (Set I and Set II) based on MHC-I and MHC-II restriction data. **5.3a)** Maximum population coverage by South Africa **5.3b)** Maximum population coverage by Europe. Set I represents epitopes LICFFTLSY, PLNPLILLY, and PIVNLFLLY from protein with Uniprot Id A7ZGR5. Set II represents epitopes SVSVFIFLF and SVFIFLFIY from protein with Uniprot Id A7ZGK4.

Table 5.4: Five most potential B-cell epitopes by combined predictions of AAP, BCPred, IEDB tools (Emini Surface Accessibility, Karplus and Schulz flexibility, Parker hydrophilicity), further filtered based on their antigenicity values (VaxiJen Score).

Table 5.5: Potential conformational B-cell epitope residues predicted using CBTOPE and DiscoTope server, along with the corresponding secondary structure conformation each residue adopts and DiscoTope scores.

Table 6.1: A list of metabolic pathways unique to *C.jejuni* against human host and the no of proteins associated with the corresponding pathway.

Table 6.2: Essential non-homologous proteins of *C. jejuni* with results of prioritization analysis (transmembrane and cellular location prediction).

Table 6.3: Essential non-homologous proteins of *C.jejuni* inferred as similar to binding partners of drugs available in Drug Bank by homology search.

Table 7.1: Characterization of essential therapeutic drug targets.

Table 7.2: Similar binding partners of drugs available in Drug Bank inferred by homology search, druggability, assayability, α -helices and pathway information on prioritized drug target proteins of *S. typhimurium* LT2.

"I may not have gone where I intended to go, but I think I have ended up where I needed to be."

~Douglas Adams~

"Perfection is not attainable, but if we chase perfection we can catch excellence."

~Vince Lombardi~

ABSTRACT

ABSTRACT

Diarrhea which is a manifestation of intestinal dysfunction is characterized by increase in stool frequency, loss of water, electrolytes, and/or nutrients. Diarrheal diseases are the second leading cause of children deaths under age five worldwide. Diarrheal diseases burden tends to be the highest in developing countries. Diarrhea is negatively associated with child growth, survival and cognitive development. Diarrhea is caused by a wide variety of organism including bacteria, viruses, parasites and helminthes with diverse pathogenesis mechanisms. Zinc supplementation as an adjunct to ORS is the recommended therapy for diarrheal management. Several medication categories are available to deal with different aspects of diarrheal diseases which include opiates, antibiotics, bile acid sequestrants and heavy metals. Due to emergence of clinically significant cases of antibiotic resistance, the problem of antibiotic resistance surfaced as a global health problem which can potentially jeopardize the achievements of modern medicine. The last novel class of antibiotics was discovered in 1987. The last three decades which essentially remain to be a discovery void era, have observed a rampant increase in antibiotic resistance with a steady decline in the number of new antibiotics approved. In India, regarded as the largest consumer of antibiotics, antibiotic use increased by 43% from 2000-2010 with 18% of the drugs being used without prescription. In containment of the global problem of antibiotic resistance, limited surveillance, irrational use of antibiotics in humans and livestock and lack of interest of pharmaceutical companies in development of new drugs and diagnostic tests are major impediments.

Single Nucleotide Polymorphisms have a vast potential to be utilized as molecular diagnostics for gene-disease or pharmacogenomics association studies linking genotype to phenotype. In this study, we first developed a database DBDiaSNP which is a comprehensive repository of mutations and resistance genes from various diarrheal pathogens to advance breakthroughs that will find applications from development of sequence based diagnostic tools to drug discovery. It contains information about 946 clinically significant mutations and 326 resistance genes compiled for pathogens such as DEC (Diarrheagenic *E. coli*), *Salmonella spp.*, *Campylobacter spp.*, *Shigella spp.*, *Clostridium difficile*, *Aeromonas spp.*, *Helicobacter pylori*, *Entamoeba histolytica*, *Vibrio cholera*, and viruses. It also includes mutations from hosts (humans and pigs) which render them either susceptible or resistant to a certain type of diarrhea. For future

translational research involving integrative biology and global health, the database offers veritable potentials, particularly for worldwide monitoring and personalized effective treatment of diarrheal pathogens.

Ciprofloxacin, a potent broad spectrum antibacterial agent is the first line of antibiotics against severe cases of Traveler's Diarrhea (TD). Several mutations in the quinolone resistance determining region (QRDR) of the structural target gene (*gyrA*) for quinolones are associated with increased quinolone resistance in vivo, which account for reduced affinity towards quinolones. To understand the molecular events underlying the mechanism of drug resistance, fully atomistic explicit-water solvated molecular dynamics simulations of wild-type, and mutant forms of *gyrA* in Enterotoxigenic *Escherichia coli* (ETEC) and *Campylobacter jejuni* complexed with ciprofloxacin were performed. From structural stability calculations, no clearly discernible conformational changes were observed. However, these mutations significantly alter *gyrA* residue interaction network and the overall pattern of global dominant motions in major distinctive domains of N-terminal regions of *gyrA*. This study offers insights into molecular events central to the *gyrA*-ciprofloxacin interaction, consequently would aid in design of more potent antibacterial agents with high ligand efficacy for treating drug resistant bacterial infections

With emergence of serious antibiotic resistance threats, a renewed search for new antibiotics and vaccine candidates becomes imperative. We have employed a range of omics and systems biology approaches to investigate the entire genome/proteome/metabolome of pathogens viz. ETEC, *C. jejuni* and *Salmonella typhimurium* to instigate the search for potential therapeutic drug targets and vaccine candidates. For vaccine candidate identification in ETEC and *C. jejuni*, comparative genomics and immunoinformatics approaches were used. The predicted epitopes were selected based on outer membrane localization, high antigenicity, population coverage, and interaction with many HLA-alleles. For drug target identification in *C. jejuni*, an integrated approach of comparative metabolic pathway analysis and subtractive genomics was utilized. Constraint based analysis of metabolic reconstruction model was exploited for identification of proteins with therapeutic potential in *S. typhimurium*. Many of the identified candidate drug targets have already been exploited as drug targets and a variety of other metabolites have not yet been targeted. Majority of the metabolites identified as potential drug targets are involved in cell

wall synthesis, fatty acid synthesis, and bacterial secretion systems which are crucial for bacterial growth and viability.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Diarrhea is a manifestation of intestinal dysfunction that results in increased stool frequency, characterized by loss of water, electrolytes, and/or nutrients. Diarrheal diseases rank second only to pneumonia as the leading cause of deaths among children under five years of age worldwide [1, 2]. Despite the fact that mortality rate of diarrheal diseases was substantially reduced by almost 31.1% from 1.8 million in 2000 to 1.3 million in 2013 [1] due to inception of many immunization programs for diarrheal disease control, the proportional mortality rate still is soaring with an estimated 2,34,570 under 5 children deaths in India and more than 2.5 million diarrheal deaths globally [2]. The highest mortality rate due to diarrheal diseases continues to be concentrated in resource poor countries. Figure 1.1 depicts the burden of diarrheal diseases across the globe.

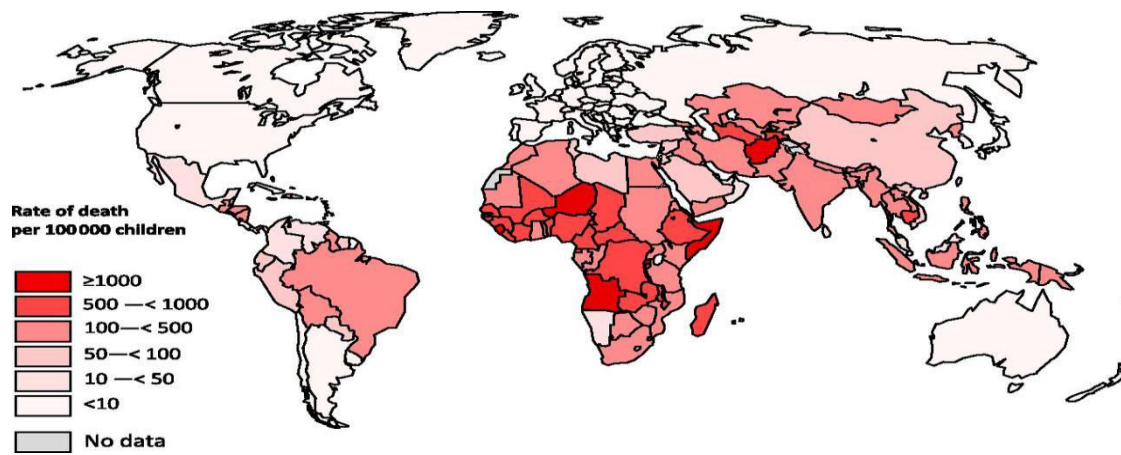


Figure 1.1: Rates of child mortality due to diarrhea².

It is estimated that 2% of diarrheal cases progress to severe diarrheal disease stage. As shown in Figure 1.2, diarrheal incidence and case-fatality ratios vary between different WHO regions and tend to be much higher in developing world than industrialized nations [3]. Diarrheal diseases are negatively associated with cognitive development, childhood growth and survival. In children affected with HIV, diarrhea is more detrimental with a reported 11 times higher mortality rate than those without diarrhea. Children below age 3 experience three diarrheal episodes annually in developing countries. In India, as shown in Figure 1.3, diarrheal diseases account for approximately 13% of all childhood deaths under age 5; additionally contributing to

another 1% of the neonatal deaths [4]. Diarrheal pediatric death toll is higher than that of AIDS, measles, and malaria combined.

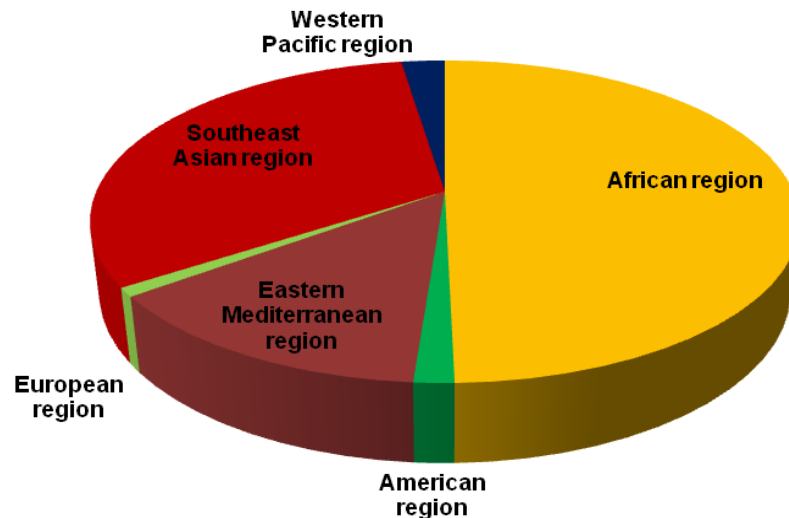


Figure 1.2: Regional burden of diarrhea per year in children aged 0–4 years, by WHO region⁵.

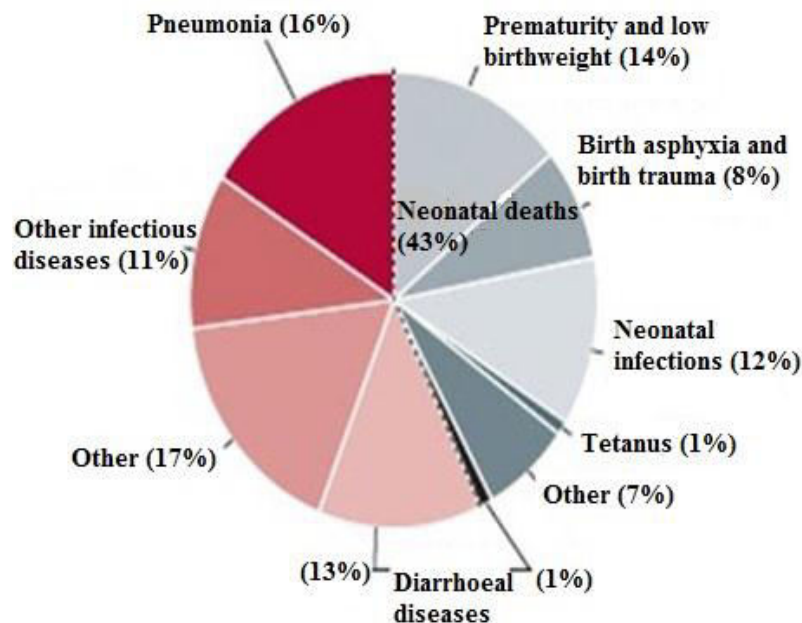


Figure 1.3: Causes of deaths in under-five children in India⁶.

Diarrhea has been classified into three categories based on clinical presentations [5].

1.) Acute Watery Diarrhea (Non-Inflammatory) which lasts for several hours or days.

2.) **Acute Bloody Diarrhea** termed as dysentery (Inflammatory).

3.) **Persistent Diarrhea** (Enteric Fever) which lasts for 14 days or longer.

Table 1.1: Classification of diarrhea based on clinical presentations

Illness	Watery Diarrhea	Dysentery	Enteric Fever
Mechanism	Noninflammatory, Enterotoxin mediated disruption of water/electrolyte secretion by gastrointestinal (GI) mucosal cells	Inflammatory with Cytotoxin mediated destruction or invasion of mucosal cells	Invasion beyond GI mucosa and dissemination systemically
Histopathology	No structural damage to GI mucosa, no inflammation	Destruction of GI mucosal cells with inflammation	
Infection Site	Small intestine (organisms generally do not penetrate GI epithelium but remain in lumen)	Large intestine (organisms actually invade but are generally limited to GI mucosa)	Distal small intestine – site of entry (disseminates systemically)
Stools Characteristics	High volume, watery	Dysentery-frequent, small volume stools containing blood and mucus	Systemic illness in which GI symptoms may not be very prominent
Presence of fecal WBCs	No	Yes (PMN)	Variable
Other clinical findings	No fever, leukocytosis; volume depletion predominates	Fever, leukocytosis; volume loss less prominent	Systemic signs and symptoms predominate-fever, headache, enlarged liver and spleen
Representative organisms	<i>Vibrio cholerae</i> , Enterotoxigenic <i>Escherichia coli</i> (ETEC), <i>Bacillus cereus</i> , <i>Clostridium perfringens</i>	<i>Shigella</i> , Enterohemorrhagic <i>Escherichia coli</i> (EHEC), <i>Campylobacter jejuni</i> , <i>Salmonella</i> , <i>Clostridium difficile</i> , <i>Entamoeba histolytica</i>	<i>Salmonella</i> , <i>Yersinia</i>

Human small and large intestine are competent to absorb approximately 10 liters of fluid per day while functioning normally [6]. As can be seen in Table 1.2, any disruptions in the normal gut physiology results in osmotic, secretory, motility or mixed diarrhea [7,8].

Table 1.2: Classification of diarrhea based on intestinal fluid absorbance capacity

Illness	Osmotic Diarrhea	Secretory Diarrhea	Motility Diarrhea
Mechanism	increase in the amount of fluid being drawn into the lumen of the bowel	dysfunction in the ability of the intestine to reabsorb fluid as it flows through the lumen	intestines functioning abnormally, motility increased or decreased
Stools Characteristics	volume one liter daily, with neutral pH	volume under one liter, acidic with loss of more potassium than sodium	low volume, liquid stool and cramping
Representative Organisms/Causative Agents	<i>Vibrio cholerae</i> , <i>E. coli</i> , <i>Campylobacter jejuni</i> , <i>Salmonella</i> , <i>Shigella</i> , <i>Clostridium difficile</i>	decreased enzymatic availability (lactose intolerance), genetic abnormality, fat malabsorption, malnutrition	bowel changes due to inflammatory bowel disease or by irritable bowel syndrome

Diarrheal symptoms are usually dependent on the causative agent and type of diarrhea, but commonly observed symptoms include frequent watery bowel movements, cramping abdominal pain, fever, abdominal tenderness, intestinal bleeding visible in the form of bloody stools, dehydration signs etc. [8]. Diarrhea is usually spread through contaminated water, food, or objects following different pathways as shown in Figure 1.4. Contaminated water poses a serious threat to human health worldwide. In a report released by World Health Organization (WHO), unsafe water supply, sanitation and hygiene accounts for 58% of diarrheal burden, which includes 3,61,000 under age five children deaths, mostly concentrated in developing countries [9].

Diarrhea is caused by a wide spectrum of infectious organisms including viruses, bacteria, parasites, and helminths. These diarrheal agents follow diverse pathogenesis mechanisms and vary in their route of transmission, inoculum size requirement and host preferences [10]. Such as, for *Shigella* a few thousand organisms are sufficient to cause infection while for *Vibrio cholerae*

millions of organisms are required. Similarly pathogens such as rotavirus have sharp host preference while *Salmonella* display broad-range host specificity. Most cases of acute gastroenteritis are viral while bacterial pathogens are discovered in foodborne illness outbreaks cases and in severe cases of bloody diarrhea [8].

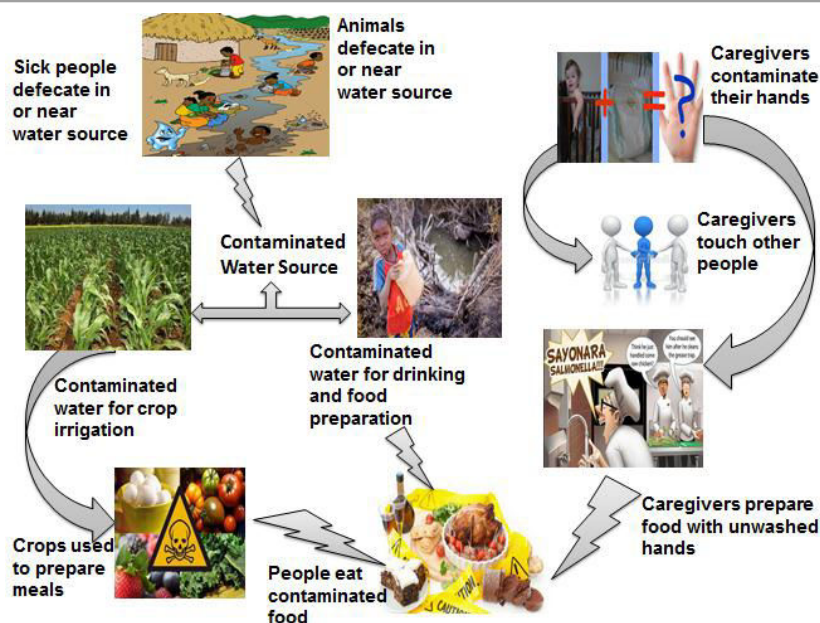


Figure 1.4: Pathways to diarrheal spread through contaminated water, food, or objects.

1.2 Diarrheal Etiology

1.2.1 Bacteria

1.2.1.1 *Escherichia coli*

E. coli is facultatively anaerobic gram negative bacillus which is the most versatile bacterial pathogen as some strains are part of normal gut flora while others cause infections by virtue of their virulence factors. Diarrheagenic *E. coli* (DEC) are divided into six distinct categories based on their pathogenesis mechanisms which include enterotoxigenic *E. coli* (ETEC), enteropathogenic *E. coli* (EPEC), enteroinvasive *E. coli* (EIEC), enterohemorrhagic *E. coli* (EHEC), enteroaggregative *E. coli* (EAEC), and diffusely adherent *E. coli* (DAEC) [11]. ETEC colonize small intestine using fimbrial adhesions and then release plasmid-associated enterotoxins heat labile (LT) or heat stable (ST) which results in secretory diarrhea [12]. ETEC is

the most common cause of traveler's diarrhea (TD) commonly spread through consumption of contaminated water. EPEC causes diarrhea by destroying microvilli in the small intestines through attaching-effacing mechanism [13] and is associated with sporadic cases and outbreaks of infection in infants and young children. EHEC causes dysentery by releasing cytotoxic Shiga Toxins (Stx-1 and Stx-2) which destroy intestinal villi [14]. It is usually spread through contaminated food and unpasteurized milk. ETEC and EPEC are the largest contributors to diarrheal incidences worldwide while EHEC is mostly concentrated in developed countries. EIEC destroys epithelial cells in the large intestines by invading them through endocytosis and results in bloody diarrhea [15]. EIEC spread is usually food-borne with no evidence of animal or environmental reservoir. EAEC colonize small intestine in an aggregative fashion [16] by using plasmid-associated fimbriae and cause persistent diarrhea. DAEC role in diarrheal diseases is poorly understood.

1.2.1.2 Vibrio species

Vibrios are curved-rod shaped gram negative bacilli which thrive in estuarine and marine environments with high salinity and temperature (up to 37°C). *V. cholerae* spreads through contaminated water and food; person-to-person transmission is unusual due to high inoculum size requirement. The symptoms range from asymptomatic colonization to fatal diarrhea rapidly after 2-3 days of bacterial ingestion which results in severe dehydration, metabolic acidosis (bicarbonate loss), hypokalemia (low potassium) are other symptoms [17]. If left untreated, the mortality rate is as high as 60%, but reduced to less than 1% through prompt replenishment of lost fluids and electrolytes [18]. *V. parahaemolyticus* which is transmitted through contaminated shellfish is another cause of mild to severe secretory diarrhea [19].

1.2.1.3 Shigella species

Shigella are rod-shaped, non-motile gram-negative bacteria which are transmitted through fecal-oral route. Shigellosis known as bacillary dysentery is characterized by an invasive infection of colonic mucosa resulting in mucus or blood in the stool [20]. Like *V. cholerae*, humans are only known reservoirs for *Shigella*. *S. sonnei*, *S. flexneri*, *S. dysenteriae*, *S. boydii* are known species where *S. sonnei* is reported more frequently in developed countries and *S. flexneri* predominantly in resource poor countries. *S. dysenteriae* infects host through shiga toxin [20] similar to that of

EHEC which damages intestinal epithelium and glomerular endothelial cells leading to kidney failure. *S. boydii* is less frequently isolated.

1.2.1.4 Campylobacter species

Campylobacter are microaerophilic gram negative bacteria with poultry as its main reservoir. The disease is spread to humans either through consumption of contaminated food or contact with infected animals. Out of thirteen species implicated in human diseases, *C. jejuni* and *C. coli* are predominantly associated with Campylobacter gastroenteritis. *C. lari* has been associated with recurrent diarrhea in children. Campylobacter damages mucosa of small and large intestines by invading into intestinal cell walls [21].

1.2.1.5 Salmonella species

Salmonella are rod-shaped facultatively anaerobic gram negative bacteria with *Salmonella enterica* and *Salmonella bongori* species. *S. enterica* includes six subspecies which include over 2500 serovars. *S. enterica* typically is classified into typhoidal (*S. typhi* and *S. paratyphi*) and non-typhoidal (*S. enteritidis* and *S. typhimurium*) species [22]. Non-typhoidal *Salmonella* species display broad host range as they are capable of infecting virtually all host species. Large inoculum is required to cause disease as ingested bacteria are killed by gastric acidity; nevertheless bacteria have been evolving to be tolerant to acidic environments of GI [23]. Non-typhoidal *Salmonella* infection in humans occurs usually through contaminated foods. Infants, elderly and immunocompromised individuals are at greater risk as small numbers of ingested bacteria are capable of disease inception.

1.2.1.6 Clostridium species

Clostridium are anaerobic spore-forming gram positive bacilli which include about 100 species of free living bacteria and pathogens. *C. perfringens* is a frequently reported cause of foodborne illness in developed countries. *C. perfringens* infects host by action of enterotoxin or β -toxin [24]. More common enterotoxin-mediated infection spreads through undercooked meat or poultry. β -toxin mediated infection results in bloody diarrhea with 50% mortality rate. *C. difficile* is commonly diagnosed to be the causative agent of nosocomial infectious diarrhea. Broad

spectrum antibiotic treatments disturb the normal gut flora and allow *C. difficile* to thrive [25]. It produces toxins which result in diarrhea.

1.2.1.7 Yersinia enterocolitica

Y. enterocolitica are rod-shaped gram negative bacteria which are motile at optimum temperature range 22-29°C [26]. It can colonize the GI tract of a variety of hosts including pigs, rodents, dogs, sheep, cattle, horses and even mammals from terrestrial and aquatic niches. Transmissions from pigs and dogs to humans have been reported. The organism usually spreads from contaminated water, meat, or milk resulting in watery or bloody diarrhea and fever.

1.2.1.8 Bacillus cereus

B. cereus is a motile gram positive bacillus which causes foodborne illness and results in severe diarrhea and gastrointestinal pain. *B. cereus* is acquired from improperly cooked food which leads to survival of bacterial endospores. It produces enterotoxins highly resistant to temperature and acidic pH. Hemolysin BL (Hbl), heat labile nonhemolytic enterotoxin (Nhe) and cytotoxin K (CytK) toxins result in diarrheal symptoms [27].

1.2.2 Viruses

Rotavirus is a double-stranded RNA virus which targets the epithelial villi of the small intestine and following infection, enterotoxin stimulates chloride secretion which results in watery diarrhea. Rotavirus is highly infectious with infectious dose between 10-100 virions and the most important cause of severe diarrheal episodes in children under age 5 (Figure 1.5). Rotavirus is responsible for approximately 22% of under age 5 deaths in India with highest mortality rates in Asian and African countries [28,29]. Noroviruses are single-stranded RNA viruses implicated in diarrhea associated with foodborne or waterborne illness. It spreads to humans through consumption of sewage-contaminated shellfish. These are highly infectious organisms with a rapid dispersal rate. They have been reported to be an important cause of diarrhea in healthcare settings and in crowded places with poor hygiene in older children and adults [30]. Astroviruses are single-stranded RNA viruses with characteristic five- or six pointed star-like surface which result in mild diarrheal episodes mostly in children. Astroviruses are responsible for 5-9% of diarrheal incidence in children [31]. Adenoviruses are unenveloped double-stranded DNA

viruses associated with gastroenteritis in children [32]. Types 40 and 41 of adenovirus are second only to rotavirus as a cause of acute diarrhea.

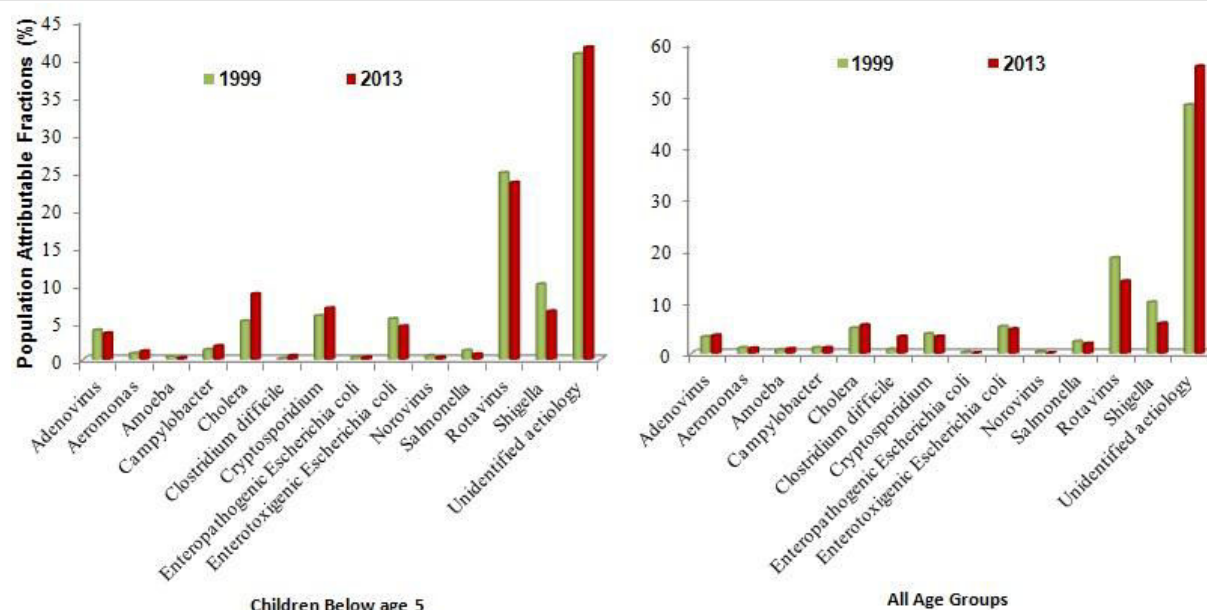


Figure 1.5: Distribution pattern of causes of diarrheal death in a) children below age 5 and b) all age groups globally from 1999-2013³.

1.2.3 Parasites

Entamoeba histolytica is an anaerobic parasitic protozoan which is mostly reported in subtropical and tropical countries with prevalence rate exceeding 50%. *E. histolytica* invades mucosal cell linings in the large intestine which results in mild diarrhea [33]. Severe invasion leads to amoebic dysentery. It spreads through contaminated food or water and can also be transmitted as a result of anal sexual activity. *E. histolytica* incidence rate is approximately 50 million which may be an underestimation as only 10%-20% of infected individuals become symptomatic. *Giardia lamblia* is a flagellated protozoan parasite which colonize small intestine of many mammalian species including cattle, sheep, dogs, birds. It is transmitted through fecal-oral route along with zoonotic transmission. *G. lamblia* can result in chronic diarrhea in immunosuppressed individuals [34]. *G. lamblia* is a major cause of epidemic childhood diarrhea with prevalence rates of 15-20% in children aged less than 10 years in the resource poor settings. *Cryptosporidium* are implicated in gastrointestinal illness which primarily involves watery

diarrhea. Major species include *C. hominis* which infects human and *C. parvum* largely infecting animals. *Cryptosporidium* infection ranges from mild to profuse watery diarrhea which can become chronic in immunocompromised individuals [34]. *Cyclospora cayetanensis* is a cause of TD which can be acquired through contaminated food also [35]. *Isospora belli* results in persistent diarrhea in AIDS patients [36]. *Enterocytozoon bieneusi* and *Encephalitozoon intestinalis* are also recognized to be diarrhea causative agents in AIDS and other immunosuppressed patients [37,38]. Mild to profuse *Trichuris* infection can also result in chronic diarrhea [39].

1.3 Diagnosis

Initial clinical evaluations focus on symptomatic diagnosis which includes assessing disease severity and degree of dehydration, identifying probable etiologic agent based on patient history and stool characteristics. Presence of visible blood in patient stool points to invasive infection due to pathogens such as *Shigella*, *C. jejuni*, *Salmonella*, or *E. histolytica* [40]. Stool cultures are necessary for *V. cholerae* identification. Incubation period, patient travel history in context to regional differences in pathogen prevalence, unusual eating situation, recent use of antimicrobials can also shed light on diarrhea epidemiology [40].

1.4 Treatment

Regardless of the etiological agent dehydration remains a critical concern. Replenishment of lost electrolytes is a critical therapy which is accomplished by intake of commercially prepared Oral Rehydration Solution (ORS) [6]. Zinc supplementation as an adjunct to ORS is recommended for diarrheal management in malnourished children and in persistent diarrhea particularly in developing countries [40]. Probiotics have been proven to be successful in reducing the length and severity of infectious diarrhea, nevertheless probiotics use may not be appropriate in resource poor countries as their effects are strain specific [40]. Depending on etiology of diarrhea, specific dietary measures can be adopted. In case of lactose intolerance as the underlying cause of diarrhea, Lactaid® substance maybe added to the diet [41]; in case of fat malabsorption, a low fat diet is recommended. Several medication categories are available to deal with different aspects of diarrheal diseases which include opiates, antibiotics, bile acid sequestrants and heavy metals. Table 1.3 lists currently available antibiotics and vaccines against

diarrheal pathogens which are either in use or under different phases of development. iOWH032 is a new drug which is under investigation for exploring its potential to be the first synthetic drug against secretory diarrhea [42].

Table 1.3 List of currently available drugs and vaccines against diarrheal pathogens.

Organism	Drugs	Vaccines		
		Live attenuated	Killed whole-cell vaccines	Subunit
<i>Shigella</i>	Ciprofloxacin, Pivmecillinam, Azithromycin	CVD1208S, WRSS1	ETEC whole cell	Chemical glycoconjugate, Recombinant glycoconjugate, Synthetic glycoconjugate, InvaplexAR, GMMA, OMV, DB Fusion, PSSP-1, 34 kDa OMP
<i>Salmonella</i>	Azithromycin, Fluoroquinolones, Ceftriaxone	JOL919, JOL1364 (C hickens)		
<i>Campylobacter</i>	Erythromycin, Azithromycin and Fluoroquinolones			
ETEC	Ciprofloxacin and Azithromycin	ACE527, ZH9	ETVAX	Dukoral®(CTB), dmLT, ST, LT patch, Fimbrial tip adhesion, CF/CS consensus peptide fused to dmLT-ST mutants, YghJ, EatA
Rotavirus		RotaTeq®, Rotarix, Rotavac		
<i>V. cholera</i>	Doxycycline and Tetracycline	CVD 103-HgR, Mutacol	Dukoral, Shanco 1	
<i>E. histolytica</i>	Metronidazole, Tinidazole, Nitroimidazole			
<i>G. lamblia</i>	Metronidazole and Ornidazole	α 1- Giardin		

Opiates such as loperamide (Imodium®) and diphenoxylate with atropine (Lomotil®) are used to treat chronic diarrhea by slowing down within the intestine passage time to allow more fluid to

be reabsorbed [43]. These medications are not to be used in case of bloody diarrhea, or in case of presence of blood and leukocytes in the stool [44]. Lomotil®) is not recommended for children below age 2. Bile acid sequestrants are used in patients who have had their gall bladder removed to neutralize the bile acid built up in the colon [43]. Cholestyramine (Questran®) is one such drug which should be used with caution due to its potential to cause adverse drug reactions (ADR) particularly in patients with renal insufficiency [6]. Heavy metals such as bismuth have been effective in TD treatment due to its anti-infective and anti-secretory properties. Pepto-Bismol® and Kaopectate® are two medications which contain bismuth subsalicylate [45] and should not be used for children below age 12. Since diarrhea is a self-limiting disease, antibiotic treatment is usually not recommended. Antibiotics are commonly required in case of TD, persistent and invasive diarrhea and *C. difficile* infection. The elderly and children are at greater risk for contacting severe diarrheal episodes. With age, GI tract changes in respect to passage time and blood flow to the gut in elderly population [46]. Children are more prone to dehydration as their metabolic rates are high while functional reserves are low [47]. Consequently, in elderly and young populations drug dosages should be reduced and time between doses should be extended to avoid any adverse effects.

1.5 Antimicrobial resistance (AMR)

AMR which is the ability of the microorganism to develop resistance to the drugs to which it was originally susceptible has emerged as a global health concern threatening the treatment of ever-widening range of infectious diseases. Antibacterial resistance (ABR) is intricate as broad range resistance mechanisms are involved affecting ever-growing bacterial spectrum which are implicated in a wide variety of diseases. Penicillin discovery in 1928 by Alexander Fleming marked the commencement of a golden era of antibiotics. Penicillin which was released for public use in 1945 was used widely in World War II. The first case of penicillin resistance emerged in 1940 [48] but clinically significant resistance cases of penicillin were reported in early 1950's. Figure 1.6 shows the timeline of antibiotic discovery and emergence of clinical resistance for various classes of antibiotics [49]. From 1940-1980, average resistance time for a drug was exceeding ten years, nevertheless in present scenario it has reduced to one year [50].

ABR is a natural phenomenon further accelerated by irrational use of antibiotics including in livestock, inadequate infection control practices, poor hygiene and lack of resistance surveillance

and monitoring systems. AMR impact on human health and health care expenditures is largely unidentified. Economic impacts of AMR to the US health system alone were estimated at US 21-34 billion dollars annually which could potentially jeopardize the achievements of modern medicine [51].

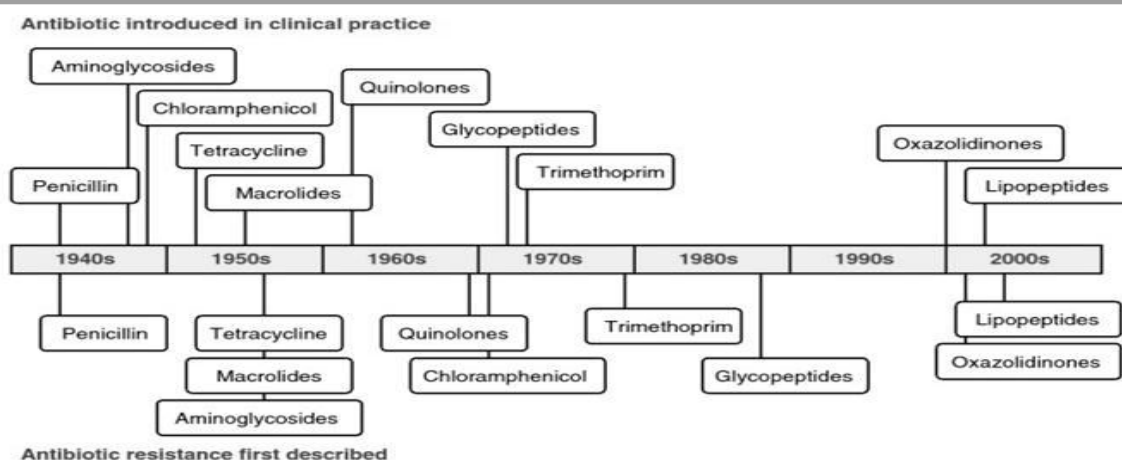


Figure 1.6: Timeline of the discovery of antibiotic classes showing the dearth of progress made in the last few decades⁵¹.

Since last three decades which apparently is a discovery void era very few antibiotics have made it to the development pipeline. A novel antibiotic class last discovered was in 1987 and antibiotics discovered thereafter are modifications of the existing antibiotics [52]. Efforts invested in research and development for antibiotics discovery have reduced significantly in last few decades. The number of new antibiotics approved is steadily decreasing [50] while the resistance among different pathogens has been escalating rapidly.

1.5.1 Antibiotic resistance mechanisms

Bacteria develop resistance through a variety of sneaky mechanisms which can be inherent or acquired. Intrinsic resistance refers to natural resistance present in bacteria owing to some genes which could generate a resistance phenotype and render it intrinsically resistant to particular antibiotic class. Acquired resistance is generated by mutations in the chromosomal genes or by acquisition of mobile genetic elements carrying antibiotic resistance; following mechanisms

shown in Figure 1.7 [53].

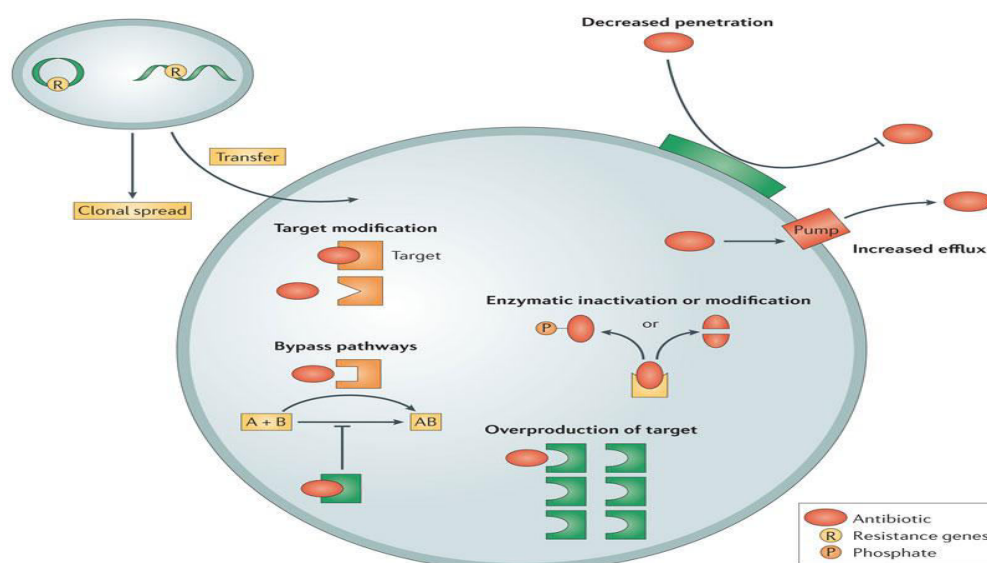


Figure 1.7: The various mechanisms of antibiotic resistance⁵⁵.

- **Inactivation of antibacterial drugs:** Bacteria can acquire genes which can either degrade the antibiotic or modify it enzymatically which render it inactive. Bacteria produce β -lactamase enzyme which renders the antibiotic inactive by hydrolyzing the β -lactam ring present on it [54].
- **Alteration of antibacterial drug target:** Some bacteria develop resistance by incorporating mutations in the antimicrobial targets which slightly modifies the drug target. The drug target can no longer bind with the antibiotic with the same affinity and exert its normal effect. *E. coli* develop resistance to quinolones by incorporating mutations in the gyraseA and B subunits [55].
- **Cell Permeability:** Bacterial cell walls have evolved non-specific porin proteins which act as passageway for antibiotics and small molecules. Bacteria hinder antibiotics entry into the cell by modifying their cell wall. *S. typhimurium* modifies lipopolysaccharide and lipid A which reduces cell permeability against polymyxin [56]
- **Efflux Pumps:** Bacteria can expel out the drug which enters the cell via efflux pumps resulting in low intracellular concentrations insufficient to cause infection. Efflux systems were originally described for tetracyclines in 1980 [57]. Efflux pumps have been

now categorized into following categories: i) the resistance-nodulation, ion-cell division (RND), ii) major facilitator (MF), iii) staphylococcal/small multidrug resistance (SMR), iv) ATP-binding cassette (ABC), and v) multidrug and toxic compound extrusion (MATE) families [58].

- **Overproduction of target:** Bacteria overproduce some enzyme substrates rendering antibiotic unable to inactivate the drug target present in abundant quantities. Enterococci develop resistance to β -lactam antibiotics by overproduction of low-affinity penicillin-binding proteins [59].
- **Bypass pathways:** Some bacteria develop high level resistance by bypassing the inactivation of enzymes through production of an alternative drug target which is resistant to a particular antibiotic. Vancomycin resistance in enterococci develops due to production of an alternative cell precursor with D-alanine-D-lactate terminus in place of normal cell precursor with D-alanine-D-alanine terminus [60].

1.5.2 Status quo of antibiotic resistance in diarrheal pathogens

In context of diarrheal pathogens, intermediate to high level of resistance has been reported globally as shown in Figure 1.8. World health leader organizations have described resistant microorganism as “nightmare bacteria” which “pose a catastrophic threat” to people around the globe as they can easily spread across continents crossing international boundaries [50]. *C. difficile* is regarded as resistant bacteria requiring urgent attention as it spreads rapidly and is naturally resistant to many antibiotics used in the treatment of other infections. It alone is responsible for 250,000 hospitalizations and 14,000 deaths per year in United States only [50]. From 2000 to 2007, 400% increase in *C. difficile* resistance has been reported. *Campylobacter* is implicated to cause estimated 1.3 million infections and 120 deaths each year in the USA. *Campylobacter* resistance to ciprofloxacin and azithromycin in USA is concerning as resistance to ciprofloxacin from 1997 (13%) to 2011 (26%) approximately doubled [50]. Non-typhoidal *Salmonella* resistance to ceftriaxone, ciprofloxacin and various other classes of antibiotics has serious implications. It causes 1.2 million infections and 450 deaths amounting to \$365 million annually in USA [50]. Resistance to third-generation cephalosporins increased from 70% to 83% in *E. coli* from 2008 to 2013, and resistance to fluoroquinolone increased from 78% to 85% [61]. A serious action is needed to tackle the drug resistant *Shigella*. *Shigella* has developed high level

resistance to first line antibiotics such as ampicillin and trimethoprim sulfamethoxazole [62]. Resistance to antibiotics like ciprofloxacin and azithromycin is growing in *Shigella* which accounts for approximately 500,000 diarrheal incidences and 40 deaths annually in USA [50]. *Shigella* has developed 50% resistance to norfloxacin and ampicillin in India [63].

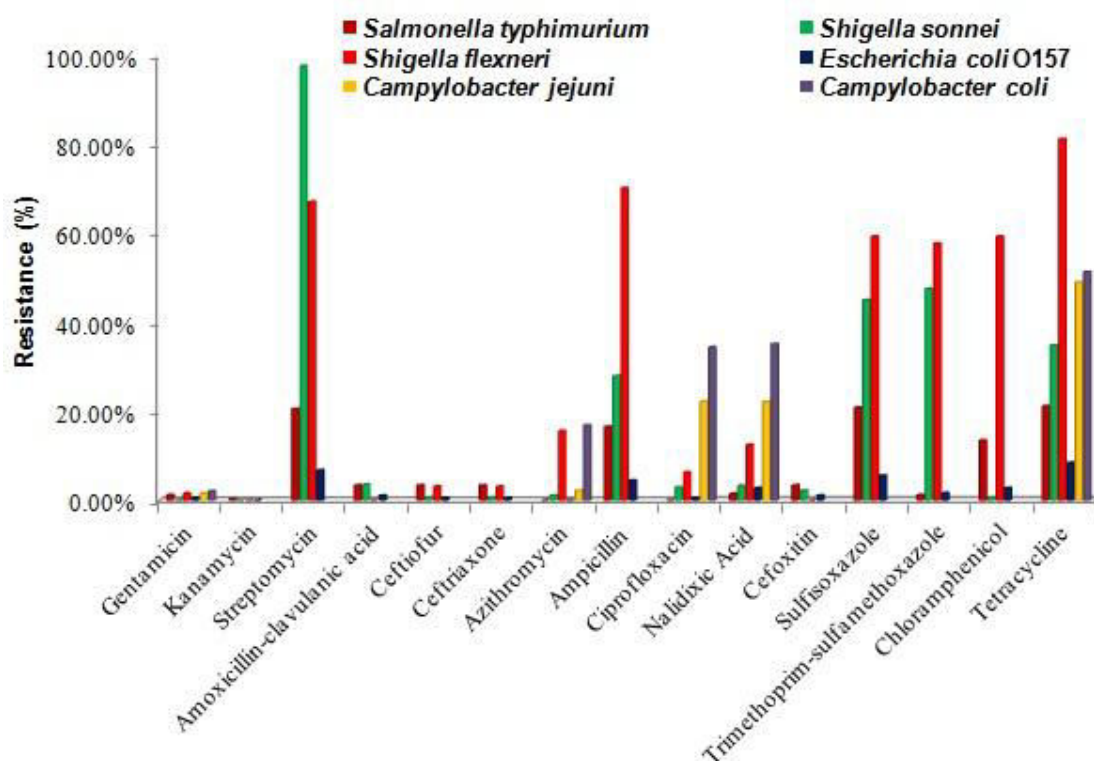


Figure 1.8: Percentage of isolates of different diarrheal pathogens resistant to antimicrobial agents, 2013⁶⁴.

1.6 Knowledge gaps

Resistance in gram negative bacteria is typically troublesome as gram negative bacteria are on the way to becoming pan-resistant while some are already there. Increasing antibiotic resistance will place us in the post-antibiotic era which will cripple the ability to treat infectious diseases jeopardizing the gains of modern medicine. In Indian context, it is not possible to clearly demarcate the antibiotic resistance trends for specific organisms due to insufficient data; nevertheless it indicates high level of resistance. Resistance is highest for the drugs in use for the longest periods of time, and for the relatively newly introduced antibiotics, resistance in India is

reported to be quite high [63]. Deaths attributable to drug resistance would be concentrated more in South-East Asia region by 2050 which is as high as 4,730,000 [63]. India regarded as the largest consumer of antibiotics in the world was reported with a whopping 43% increase in antibiotic use from 2000 to 2010 [64]; where 18% of the drugs were used without prescription [65]. Indian Council of Medical Research (ICMR) launched the Anti-Microbial Resistance Surveillance and Research Network (AMRSN) in 2013 with its seven nodes focusing on diarrhea, enteric fever, sepsis, gram-positive bacteria, fungal infections and respiratory pathogens. AMRSN in its insurgency has been limited to small scale efforts, far behind its intended goal of being established as a permanent surveillance system for antibiotic resistance. The knowledge gaps in antibiotic resistance exist in form of

- i) Limited international, national and state surveillance to monitor urgent and emerging antibiotic resistant pathogens
- ii) No scaling of antibiotic use in humans and livestock
- iii) Lack of antibiotic stewardship
- iv) Less effort invested in development of new drug and diagnostic tests.

We have attempted to contribute to tracking and benchmarking antibiotic resistance and developing new antibiotics to keep pace with ever-evolving resistant bacteria with the motivation to address aforementioned gaps in tackling the global threat of antibiotic resistance.

1.7 Objectives

This research study was undertaken to study the antibiotic resistance mechanism in diarrheal pathogens and to propose potential drug targets and vaccine candidates against them using *in silico* approaches. The objective of this study is to develop a database which can unify the molecular data of mutations and resistance genes of diarrheal pathogens curated from published literature. SNPs deposited in the database will be studied at structural level wherever possible to derive a possible explanation for genotype to phenotype correlation where genotype represents SNPs in the drug target proteins and phenotype represents change in Minimum Inhibitory Concentration (MIC) values. Finally, integrated omics approaches such as immunoinformatics, comparative genomics, subtractive genomics and metabolic modeling will be applied aimed at

predicting some potential therapeutic drug targets and vaccine candidates. The specific objectives which are being addressed are:

- **To develop a database of mutations and resistance genes from various diarrheal pathogens**
- **To study the structural signatures of the pathogen mutations deposited in the database and their consequences on drug binding landscape**
- ***in silico* prediction of potential therapeutic drug targets and vaccine candidates against a wide variety of diarrheal pathogens**

Acquisition of mutations/polymorphism/SNP (Single Nucleotide Polymorphism) is a major force underlying bacterial evolution which provides a platform for natural evolution. Mutations emerge and accumulate in bacterial pathogens at a rapid rate as they are of haploid nature for majority of their genes and have typically short generation times. High mutation rates in bacteria render significant phenotypic changes in a real-time manner. Although horizontal gene transfer between bacterial pathogens is an important contributor to antibiotic resistance, however mutational resistance plays a critical role in continuous bacterial evolution and has clinical significance [66]. SNPs proved to be potential biomarkers due to their genome-wide availability in virtually all organisms and being large in numbers. SNPs find applications and have been extensively utilized in a wide range of areas such as human forensics [67] and diagnostics [68], aquaculture [69], marker assisted-breeding of dairy cattle [70], crop improvement [71], conservation [72], and resource management in fisheries [73]. SNPs are exploited in genetic mapping to decipher breeding pedigree, in evolutionary studies to identify genomic divergence aimed at explicating the role of speciation in evolution and in genome wide association studies to correlate genotypic variations with the phenotypic traits [74].

With a rapid emergence of serious antibiotic resistance threats, there is a pressing need to monitor the spread of antibiotic resistance. Molecular information pertaining to antibiotic resistance associated mutations has substantially increased in last few decades which can bequeath an understanding into spread and mechanism of antibiotic resistance. Nevertheless, mining such information is deeply hindered due to lack of data warehouses which can integrate the molecular data. The identification and cataloguing of SNPs implicated in antibiotic resistance and resistance genes in a public repository proves to be a useful resource for various

purposes in both basic and applied research. Treatment of diarrheal diseases is heavily compromised due to a wide variety of etiological agents involved which is further compounded by a rampant increase in the antibiotic resistance in diarrheal pathogens. Large scale studies have been undertaken globally to provide a catalog of antibiotic resistance associated mutations from different pathogens such as MTCID, CRAD, ARDB etc. [75-77] There have been no efforts to curate the experimental SNPs reported to be associated with antibiotic resistance in diarrheal pathogens which is scattered all over the web. We have attempted to build a comprehensive repository of resistance genes and antibiotic resistance associated mutations curated from published literature which is discussed in detail in Chapter 2.

SNPs have important functional consequences such as in signal transduction [78], in gene regulation [79], in maintaining cell structural integrity [80] and in drug resistance by affecting drug target proteins such as G-protein coupled receptors, enzymes and ion channels [81-83]. Effects of SNPs on gene expression, mRNA stability and protein structure have important reverberations on drug response [84]. Evaluating the effect of SNPs in the context of protein three dimensional structures (3D) can yield valuable insights into antibiotic resistance mechanism by shedding light on drug-target interaction and genotype-phenotype association [85]. The efficient utilization of SNP based web resources involves i) understanding how individual genetic variations are decoded into structural modifications in protein drug targets which get translated into differential therapeutic drug responses and ii) and incorporating modifications into existing antibiotics based on how these variations alter drug response pathways. We have performed molecular modeling and molecular dynamics simulation studies to understand the effect of mutations in gyrase A subunit on quinolone affinity in ETEC and *C. jejuni* which is described in detail in Chapter 3.

Due to ever-growing antibiotic resistance, economic constraints, limited time frame for antibiotic use before emergence of drug resistance and unfavorable regulatory factors pertaining to antibiotic approval and rigorous criteria of minimum adverse side effects, pharmaceutical industries have abridged their investment in antibiotic development ultimately shrinking the therapeutic arsenal in the battle against infectious diseases. With ever-widening range of microbes being resistant to existing antibiotics, bioterrorism threat, re-emergence of previously fatal and emergence of new infectious diseases, there remains a dire need to reinvigorate the

antibiotic pipeline. We have utilized the power of omics approaches to identify potential drug targets and vaccine candidates in various diarrheal pathogens which is discussed in Chapters 4, 5, 6, and 7.

References:

1. GBD. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. In: *The Lancet*. (Elsevier, 2015) 117-171.
2. WHO. Diarrhoeal disease. (World Health Organization, 2016)
3. Walker CLF, Rudan I, Liu L *et al*. Global burden of childhood pneumonia and diarrhoea. *The Lancet*, 381(9875), 1405-1416 (2013).
4. Collaborators TMDS. Causes of neonatal and child mortality in India: a nationally representative mortality survey. In: *The Lancet*. (Elsevier, 2010) 1853-1860.
5. Bushen OY, Guerrant RL. Acute infectious diarrhea. Approach and Management in the emergency department. . *Topics in Emergency Medicine*, 25(2), 139-149 (2003).
6. Kent AJ, Banks MR. Pharmacological Management of Diarrhea. *Gastroenterology Clinics of North America*, 39(3), 495-507 (2010).
7. Sisson V. Types of Diarrhea and Management Strategies. (PharmCon, Inc, 2011)
8. Yin M. Infectious Diarrhea. www.columbia.edu
9. WHO. Preventing diarrhoea through better water, sanitation and hygiene: exposures and impacts in low- and middle-income countries. (World Health Organization, France, 2014)
10. Keusch G, Fontaine O, Bhargava A *et al*. *Diarrheal Diseases*. In: Disease Control Priorities in Developing Countries. (Washington, 2006).
11. Nataro JP, Kaper JB. Diarrheagenic Escherichia coli. *Clinical Microbiology Reviews*, 11(1), 142-201 (1998).
12. Sears CL, Kaper JB. Enteric bacterial toxins: mechanisms of action and linkage to intestinal secretion. *Microbiological Reviews*, 60(1), 167-215 (1996).
13. Jerse AE, Yu J, Tall BD, Kaper JB. A genetic locus of enteropathogenic Escherichia coli necessary for the production of attaching and effacing lesions on tissue culture cells. *Proceedings of the National Academy of Sciences of the United States of America*, 87(20), 7839-7843 (1990).
14. O'Brien A, Lively T, Chen M, Rothman S, Formal S. Escherichia coli O157:H7 strains associated with haemorrhagic colitis in the United States produce a Shigella dysenteriae 1 (SHIGA) like cytotoxin. *Lancet*, 1(8326 Pt 1), 702 (1983).
15. Small PL, Falkow S. Identification of regions on a 230-kilobase plasmid from enteroinvasive Escherichia coli that are required for entry into HEP-2 cells. *Infection and Immunity*, 56(1), 225-229 (1988).
16. Nataro JP, Hicks S, Phillips AD, Vial PA, Sears CL. T84 cells in culture as a model for enteroaggregative Escherichia coli pathogenesis. *Infection and Immunity*, 64(11), 4761-4768 (1996).

17. Faruque SM, Albert MJ, Mekalanos JJ. Epidemiology, Genetics, and Ecology of Toxigenic *Vibrio cholerae*. *Microbiology and Molecular Biology Reviews*, 62(4), 1301-1314 (1998).
18. Sack DA, Sack RB, Nair GB, Siddique AK. Cholera. *The Lancet*, 363(9404), 223-233 (2004).
19. Finkelstein R (ed.) *Chapter 24 Cholera, Vibrio cholerae O1 and O139, and Other Pathogenic Vibrios* (Galveston, 1996).
20. Schroeder GN, Hilbi H. Molecular Pathogenesis of *Shigella* spp.: Controlling Host Cell Signaling, Invasion, and Death by Type III Secretion. *Clinical Microbiology Reviews*, 21(1), 134-156 (2008).
21. Zilbauer M, Dorrell N, Wren BW, Bajaj-Elliott M. *Campylobacter jejuni*-mediated disease pathogenesis: an update. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 102(2), 123-129 (2008).
22. Santos RL, Zhang S, Tsolis ReM, Kingsley RA, Garry Adams L, Baumler AJ. Animal models of *Salmonella* infections: enteritis versus typhoid fever. *Microbes and Infection*, 3(14-15), 1335-1344 (2001).
23. Garcia-del PF, Foster JW, Finlay BB. Role of acid tolerance response genes in *Salmonella typhimurium* virulence. *Infection and Immunity*, 61(10), 4489-4492 (1993).
24. Brett MM, Rodhouse JC, Donovan TJ, Tebbutt GM, Hutchinson DN. Detection of *Clostridium perfringens* and its enterotoxin in cases of sporadic diarrhoea. *Journal of Clinical Pathology*, 45(7), 609-611 (1992).
25. Garg S, Mirza YR, Girotra M *et al.* Epidemiology of *Clostridium difficile*-Associated Disease (CDAD): A Shift from Hospital-Acquired Infection to Long-Term Care Facility-Based Infection. *Digestive Diseases and Sciences*, 58(12), 3407-3412 (2013).
26. Fabrega A, Vila J. *Yersinia enterocolitica*: Pathogenesis, virulence and antimicrobial resistance. *Enfermedades Infecciosas y Microbiologia Clinica*, 30(1), 24-32 (2012).
27. Kotiranta A, Lounatmaa K, Haapasalo M. Epidemiology and pathogenesis of *Bacillus cereus* infections. *Microbes and Infection*, 2(2), 189-198 (2000).
28. WHO. Estimates of disease burden and cost-effectiveness. (World Health Organization, 2015)
29. WHO. Estimated rotavirus deaths for children under 5 years of age: 2008, 453 000. (World Health Organization, 2012)
30. Goodgame R. Norovirus gastroenteritis. *Curr Gastroenterol Rep*, 8(5), 401-408 (2006).
31. Monroe S, Holmes J, Belliot G. Molecular epidemiology of human astroviruses. *Novartis Found Symp.*, 238, 237-245 (2001).
32. Dennehy P. Viral gastroenteritis in children. *Pediatr Infect Dis J.*, 30(1), 63-64 (2011).
33. Rossignol J-F, Ayoub A, Ayers Marc S. Treatment of Diarrhea Caused by *Giardia intestinalis* and *Entamoeba histolytica* or *E. dispar*: A Randomized, Double-Blind, Placebo-Controlled Study of Nitazoxanide. *Journal of Infectious Diseases*, 184(3), 381-384 (2001).
34. Huang DB, White AC. An Updated Review on *Cryptosporidium* and *Giardia*. *Gastroenterology Clinics of North America*, 35(2), 291-314 (2006).
35. Kansouzidou A, Charitidou C, Varnis T, Vavatsi N, Kamaria F. *Cyclospora cayetanensis* in a Patient with Travelers' Diarrhea: Case Report and Review. *Journal of Travel Medicine*, 11(1), 61-63 (2004).

-
36. Velasquez JN, Osvaldo GA, Di Risio C *et al.* Molecular characterization of *Cystoisospora belli* and unizuite tissue cyst in patients with Acquired Immunodeficiency Syndrome. *Parasitology*, 138(03), 279-286 (2011).
 37. Rabodonirina M, Bertocchi M, Desportes-Livage I *et al.* Enterocytozoon bienueusi as a Cause of Chronic Diarrhea in a Heart-Lung Transplant Recipient Who Was Seronegative for Human Immunodeficiency Virus. *Clinical Infectious Diseases*, 23(1), 114-117 (1996).
 38. Raynaud L, Delbac F, Broussole V *et al.* Identification of Encephalitozoon intestinalis in Travelers with Chronic Diarrhea by Specific PCR Amplification. *Journal of Clinical Microbiology*, 36(1), 37-40 (1998).
 39. Diniz-Santos D, Jambeiro J, Mascarenhas R, Silva L. Massive Trichuris trichiura infection as a cause of chronic bloody diarrhea in a child. *J Trop Pediatr.*, 52(1), 66-68 (2006).
 40. WGO. Acute Diarrhea in Adults and Children A Global Perspective. *J Clin Gastroenterol*, 47(1) (2013).
 41. Marchiondo K. Lactose intolerance: a nursing perspective. *Medsurg Nurs.*, 81(1), 9-15 (2009).
 42. Schwertschlag U, Kumar A, Kochhar S *et al.* Mo1675 Pharmacokinetics and Tolerability of iOWH032, an Inhibitor of the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Chloride Channel, in Normal Volunteers and Cholera Patients. *Gastroenterology*, 146(5), S-633 (2014).
 43. Bliss DZ, Doughty DB, Heitkemper MM (eds.) *Pathology and management of bowel dysfunction* (Elsevier, 2006).
 44. Bushen OY, Guerrant RL. Acute infectious diarrhea. Approach and Management in the emergency department. *Topics in Emergency Medicine*, 25(2), 139-149 (2003).
 45. Turkoski BB, Lance BR, Bonfiglio MF (eds.) *Drug information handbook For advanced practice nursing* (Ohio, 2007).
 46. Zarowitz BJ. Pharmacologic consideration of commonly used gastrointestinal Drugs in the elderly. *Gastroenterolgy clinics of North America*, 38(3), 457-562 (2009).
 47. Burpee T, Duggan C (eds.) *Chapter 57: Diarrheal diseases*. In: Nutrition in Pediatrics: Basic Science & Clinical Applications, 631(2008).
 48. Abraham EP, Chain E. An Enzyme from Bacteria able to Destroy Penicillin. *Nature*, 146, 837-837 (1940).
 49. Song YG. The History of Antimicrobial Drug Development and the Current Situation. *Infection & Chemotherapy*, 44(4), 263-268 (2012).
 50. CDC. Antibiotic Resistance Threats in the United States, 2013. (Centers for Disease Control and Prevention, 2013)
 51. WHO. Antimicrobial resistance: global report on surveillance. (World Health Organization, 2014)
 52. Silver LL. Challenges of Antibacterial Discovery. *Clinical Microbiology Reviews*, 24(1), 71-109 (2011).
 53. Lewis K. Platforms for antibiotic discovery. *Nature Reviews | Drug Discovery*, 12(371-387) (2013).
 54. Drawz SM, Bonomo RA. Three Decades of β -Lactamase Inhibitors. *Clinical Microbiology Reviews*, 23(1), 160-201 (2010).

55. Karczmarczyk M, Martins M, Quinn T, Leonard N, Fanning Sa. Mechanisms of Fluoroquinolone Resistance in *Escherichia coli* Isolates from Food-Producing Animals. *Applied and Environmental Microbiology*, 77(20), 7113-7120).
56. Falagas ME, Kasiakou SK, Saravolatz LD. Colistin: The Revival of Polymyxins for the Management of Multidrug-Resistant Gram-Negative Bacterial Infections. *Clinical Infectious Diseases*, 40(9), 1333-1341 (2005).
57. McMurry L, Petrucci RE, Levy SB. Active efflux of tetracycline encoded by four genetically different tetracycline resistance determinants in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 77(7), 3974-3977 (1980).
58. Li XZ, Nikaido H. Efflux-Mediated Drug Resistance in Bacteria: an Update. *Drugs*, 69(12), 1555-1623 (2009).
59. Huycke M, Sahm D, Gilmore M. Multiple-drug resistant enterococci: the nature of the problem and an agenda for the future. *Emerg Infect Dis*, 4, 239-249 (1998).
60. Leclercq R, Courvalin P. Resistance to glycopeptides in enterococci. *Clin Infect Dis*, 24, 545-554 (1997).
61. Laxminarayan R, Chaudhury RR. Antibiotic Resistance in India: Drivers and Opportunities for Action. *PLoS Med*, 13(3), e1001974 (2016).
62. CDC. National Antimicrobial Resistance Monitoring System: Enteric Bacteria. Human Isolates Final Report, 2013. Services, USDoHaH (Centers for Disease Control and Prevention, Atlanta, Georgia, 2015)
63. Ganguly N, Arora N, Chandy S *et al.* Rationalizing antibiotic use to limit antibiotic resistance in India. *The Indian Journal of Medical Research*, 134(3), 281-294 (2011).
64. Gelband H, Miller-Petrie M, Pant S *et al.* The State of the World's Antibiotics, 2015. (Centre for Disease Dynamics, Economics and Policy, 2015)
65. Morgan DJ, Okeke IN, Laxminarayan R, Perencevich EN, Weisenberg S. Non-prescription antimicrobial use worldwide: a systematic review. *The Lancet Infectious Diseases*, 11(9), 692-701 (2011).
66. Woodford N, Ellington MJ. The emergence of antibiotic resistance by mutation. *Clinical Microbiology and Infection*, 13(1), 5-18 (2007).
67. Brenner CH, Weir BS. Issues and strategies in the DNA identification of World Trade Center victims. *Theoretical Population Biology*, 63(3), 173-178 (2003).
68. McCarthy MI, Abecasis GR, Cardon LR *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5), 356-369 (2008).
69. Liu ZJ, Cordes JF. DNA marker technologies and their applications in aquaculture genetics. *Aquaculture*, 238(1&2), 1-37 (2004).
70. Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 123(4), 218-223 (2006).
71. Yu H, Xie W, Wang J *et al.* Gains in QTL Detection Using an Ultra-High Density SNP Map Based on Population Sequencing Relative to Traditional RFLP/SSR Markers. *PLoS ONE*, 6(3), e17595 (2011).

-
72. Seddon JM, Parker HG, Ostrander EA, Ellegren H. SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. *Molecular Ecology*, 14(2), 503-511 (2005).
 73. Smith CT, Elfstrom CM, Seeb LW, Seeb JE. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. *Molecular Ecology*, 14(13), 4193-4203 (2005).
 74. Kumar S, Banks TW, Cloutier S. SNP Discovery through Next-Generation Sequencing and Its Applications. *International Journal of Plant Genomics*, 2012, 15 (2012).
 75. Bharti R, Das R, Sharma P, Katoch K, Bhattacharya A. MTCID: a database of genetic polymorphisms in clinical isolates of *Mycobacterium tuberculosis*. *Tuberculosis (Edinburgh, Scotland)*, 92(2), 166-172 (2012).
 76. Liu B, Pop M. ARDB--Antibiotic Resistance Genes Database. *Nucleic acids research*, 37(Database issue), D443-447 (2009).
 77. McArthur AG, Wagelchner N, Nizam F *et al.* The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy*, 57(7), 3348-3357 (2013).
 78. Neidhardt J, Barthelmes D, Farahmand F, Fleischhauer JC, Berger W. Different Amino Acid Substitutions at the Same Position in Rhodopsin Lead to Distinct Phenotypes. *Investigative Ophthalmology & Visual Science*, 47(4), 1630-1635 (2006).
 79. Savage DB, Tan GD, Acerini CL *et al.* Human Metabolic Syndrome Resulting From Dominant-Negative Mutations in the Nuclear Receptor Peroxisome Proliferator-Activated Receptor. *Diabetes*, 52(4), 910-917 (2003).
 80. Hoefele J, Mayer K, Scholz M, Klein H-G. Novel PKD1 and PKD2 mutations in autosomal dominant polycystic kidney disease (ADPKD). *Nephrology Dialysis Transplantation*, 26(7), 2181-2188 (2011).
 81. Kooffreh ME, Anumudu CI, Duke R, Okpako EC, Kumar PL. Angiotensin II type 1 receptor A1166C gene polymorphism and essential hypertension in Calabar and Uyo cities, Nigeria. *Indian Journal of Human Genetics*, 19(2), 213-218 (2013).
 82. Sim E, Lack N, Wang CJ *et al.* Arylamine N-acetyltransferases: Structural and functional implications of polymorphisms. *Toxicology*, 254(3), 170-183 (2008).
 83. Tristani-Firouzi M, Sanguinetti Michael C. Structural determinants and biophysical properties of HERG and KCNQ1 channel gating. *Journal of Molecular and Cellular Cardiology*, 35(1), 27-35 (2003).
 84. Daly Ann K. Pharmacogenetics and human genetic polymorphisms. *Biochemical Journal*, 429(3), 435-449 (2010).
 85. Weigelt J. Structural genomics-Impact on biomedicine and drug discovery. *Experimental Cell Research*, 316(8), 1332-1338 (2010).

CHAPTER 2

DBDiaSNP: An Open-Source Knowledgebase of Genetic Polymorphisms and Resistance Genes related to Diarrheal Pathogens

2.1 Introduction

Diarrheal diseases continue to be a significant cause of morbidity and mortality, especially among children under five in developing countries. Oral Rehydration Therapy (ORT) is the cornerstone of treatment as recommended by the WHO [1], but is not sufficient therapy for some cases of bloody diarrhea, severe dehydration, intractable vomiting, monosaccharide mal-absorption. Emergence of antimicrobial resistance, especially from the multidrug resistant microbes, is challenging the global prospects for fighting the diarrheal pathogens. Acquired resistance that results from mutation or acquisition of foreign resistance genes is a major player in the resistance against antibiotics [2]. Mutations and the resistance genes in the pathogen genome affect the drug response by various mechanisms such as inactivation of antimicrobial agents, modification of the antimicrobial target, altering membrane permeability, efflux pumps etc. [3]; making it crucial to understand this trilateral relationship between disease and genome variations and resistance genes.

Single Nucleotide Polymorphisms (SNPs) are the unique genetic differences between individuals that contribute significantly in identification and forensics [4], mapping and genome wide association studies to complex diseases [5], to predict specific genetic traits [6], to classify patients in clinical trials [7]. SNPs underlie differences in susceptibility to or protection from a host of diseases thereby proving its great potential in personalized medicine [8]. Therefore, a database of the variations in a group of organisms is not only useful for understanding genotype-phenotype relationship but also in clinical applications. The phenomenon of gene acquisition is mediated by transformation, transduction and conjugation. Plasmids, transposons, and integrons and gene cassettes are the platforms on which these resistance genes are assembled. They allow bacterial strains to expand their niches to the areas that were previously denied [9].

Large scale studies have been undertaken throughout the world to identify the antibiotic resistance associated mutations which are devoted to some specific organism like *Mycobacterium tuberculosis* (MTCID, MGDD, MTBC, GMTV) or which are generalized and host mutations or resistance data from different organisms (CRAD, ARDB, CREAM) [10-17].

The information about pathogens which are implicated in other disorders including diarrhea, is not disease specific while our database pertains to clinical isolates of diarrheal pathogens.

Till date, there has been no single resource available that provides information about genomic variations and resistance genes among different diarrheal pathogens; readily available to be used in research and development. We have developed DBDiaSNP in an attempt to provide a single platform for storage and retrieval of large amount of polymorphism and resistance gene data, particularly which is derived from clinical isolates of diarrhea. It complements our other database dbDiarrhea [18] which is a repertoire of pathogen proteins and vaccine antigens implicated in the pathogenesis of diarrhea. This database provides information about prevalence of resistance genes or mutations in a specific geographical region; increasing our knowledge on which mutations or resistance genes are common in certain bacteria or geographical locations. Mutation data in such databases can be used to screen newly sequenced strains for presence of any previously characterized resistance associated mutation; ultimately paving a way for information based targeted therapies. By mapping these SNPs at structural level, we can glean insights into the molecular level details of antibiotic resistance due to mutations. We can study how these mutations affect the binding of antibiotics to the target proteins. Such studies offer us with the possibility of mitigating the ever growing problem of antibiotic resistance by incorporating necessary modifications in already existing antibiotics.

2.2 Methods

2.2.1 Database system implementation

DBDiaSNP has been designed on MySQL 5.1 that provides multi-layered server design with fast execution times. It is implemented by using three-tier architecture. DBDiaSNP was developed on an IIS server (7.0) in a Windows operating system. The web pages were written using PHP, HTML language, Cascading Style Sheets (CSS) and JavaScript. The IIS server handles queries from web clients through PHP scripts to perform searches which allow researchers to obtain SNP information. An overview of the DBDiaSNP architecture is given in Figure 2.1.a. As shown in Figure 2.1.b, the database is a comprehensive resource of drug-resistance mutations and resistance genes in diarrheal pathogens and hosts. We conducted a systematic review to identify

antibiotic resistance mutations and resistance genes reported in the literature to be included in the database. The variations are categorized into SNPs, substitutions, frameshift mutations, deletions, insertions. For each mutation, the database provides complete codon change at nucleotide or amino acid level. Resistance genes reported in different pathogens can also be browsed and are categorized into Plasmid mediated, Transposons mediated, Integrons and gene Cassettes mediated resistance modes according to its subcellular location. Table 2.1, 2.2 and 2.3 show the database statistics which houses 946 mutations with clinical relevance and 326 resistance genes associated with diarrheal pathogens (viruses, protozoans and bacterial species) and hosts. Table 2.4 shows the breakup of source of SNPs deposited in database. In pathogens, mutations and resistance genes are responsible for antibiotic resistance. The host mutations are the ones which render them susceptible or resistant to a particular type of diarrhea or to diseases where diarrhea is a predominant symptom.

2.2.2 SNP and resistance data search

DBDiaSNP allows users to search for SNP and resistance gene information based on organism of interest which includes both host and pathogens. Related literature citations can also be retrieved based on the resistance mechanism adopted. The search results provide related SNP and resistance gene data and their annotations. The information on corresponding protein, PDB identifiers, function etc. associated with each of the SNPs and resistance gene has been extracted from different web resources dbSNP, OMIM, ClinVar, PDB [19-22]; thus providing a comprehensive compilation of the SNPs and resistance genes reported in the clinical isolates of various diarrheal pathogens.

2.3 Results and discussion

DBDiaSNP presents a database entirely devoted to drug-resistance mutations and resistance genes in diarrheal pathogens. We hope it will open new avenues to the discoveries that will have applications from diagnostics to drug discovery.

2.3.1 Development of SNP-based diagnostics

There is pressing need to develop tools that can identify the antibiotic-resistant pathogens rapidly and with high accuracy. SNPs based diagnostic methods detect mutations associated with a particular disease, but such tools extensively rely on information about the mutations and their relative frequency [23]. As more sequencing data are made available, this database project, we believe will pave a path for the currently available tools to detect resistance to the antibiotics.

Table 2.1: SNPs statistics along with genes:

Species wise distribution in pathogens

Organism	#SNPs (Genes)
Viruses	221 (10)
<i>Escherichia coli</i>	91 (15)
<i>Entamoeba histolytica</i>	11 (1)
<i>Campylobacter</i> spp.	98 (4)
<i>Shigella</i> spp.	13 (3)
<i>Vibrio cholerae</i>	16 (1)
<i>Aeromonas caviae</i>	4 (2)
<i>Clostridium difficile</i>	46 (4)
<i>Helicobacter pylori</i>	71 (3)
<i>Salmonella</i> spp.	19 (3)
Total	590 (46)

Table 2.2: Resistance Gene statistics:

Species wise distribution in pathogens

Organism	#Resistance Genes
<i>Escherichia coli</i>	68
<i>Enterococcus</i> spp.	10
<i>Campylobacter</i> spp.	10
<i>Shigella</i> spp.	58
<i>Vibrio</i> spp.	60
<i>Aeromonas</i> spp.	26
<i>Clostridium</i> spp.	10
<i>Klebsiella</i> spp.	16
<i>Salmonella</i> spp.	68
Total	326

Table 2.3: SNPs statistics along with genes: Species wise distribution in hosts

Organism	#SNPs (Genes)
Humans	291 (57)
Porcine	65 (3)
Total	356 (60)

Table 2.4: Breakup of Source of SNPs in database

Source	# SNPs
Pubmed	590
OMIM	229
dbSNP	102
clinVar	25

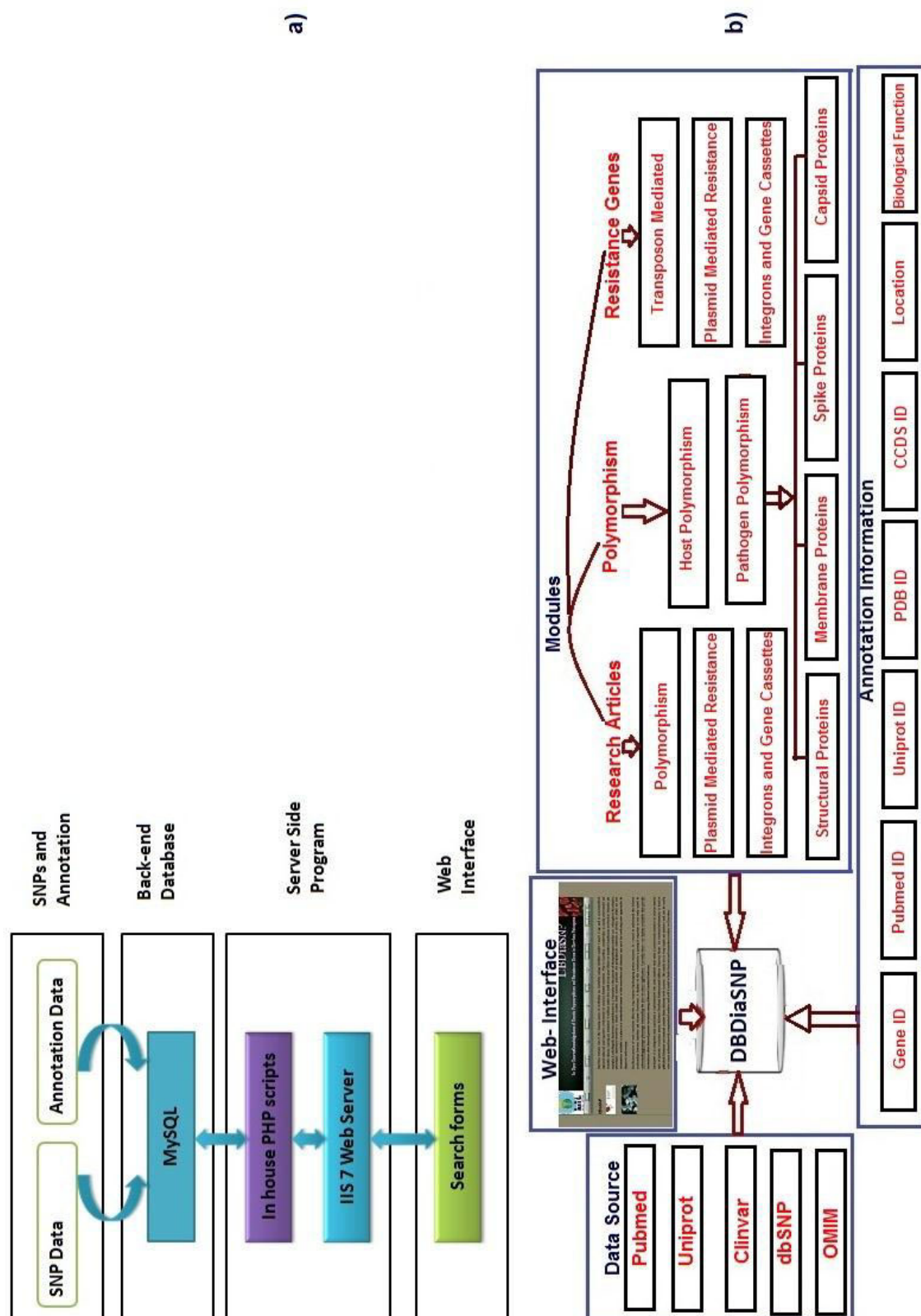


Figure 2.1: DBDiaSNP architecture overview. **a)** Standard three-tier architecture of DBDiaSNP; **b)** Organizational structure of the DBDiaSNP. The data were compiled from existing literature and online databases Uniprot, ClinVar, dbSNP, and OMIM, and enriched with annotation information such as PDB ID, Gene ID, Pubmed ID, Biological function, and CCDS ID from online resources. The database includes three modules: **1)** Polymorphism data further categorized into Host and Pathogen Polymorphism; **2)** Resistance Gene data further categorized into Plasmid mediated, Integrons, and Gene Cassettes, and Transposon Mediated; **3)** Research articles relevant to each category.

2.3.2 SNPs monitoring in different geographical regions

Different analogs of the same drug interact differently with the targets. Mutations arising as a result of treatment with different drug analogs may help explicate the geographic differences in drug effectiveness [24]. Therefore, we have mentioned the geographic locations of antibiotic resistance studies in our database. Availability of data on frequency of mutations through sequencing projects will lead to better understanding of regional differences in drug resistance.

2.3.3 Drug discovery

SNPs can have a major influence on response to pharmacotherapy [25]. The database will aid in lead discovery against antibiotic resistant pathogens by mapping the mutations onto the protein structure; providing a better understanding of molecular mechanism of drug-resistance. Additionally such analysis may suggest drug modifications to combat emerging drug resistance in diarrheal pathogens.

We have created an interactive web interface to retrieve information about mutations and resistance genes in diarrheal pathogens associated with drug resistance. Currently there is information from major diarrheal pathogens and hosts. To get a comprehensive list of variations, particularly substitutions and in-dels; one can choose Host Polymorphism or Pathogen Polymorphisms. As shown in Figure 2.2 for both pathogen and host mutations, after submission, the output will show all the SNPs along with annotation of genes corresponding to these SNPs.

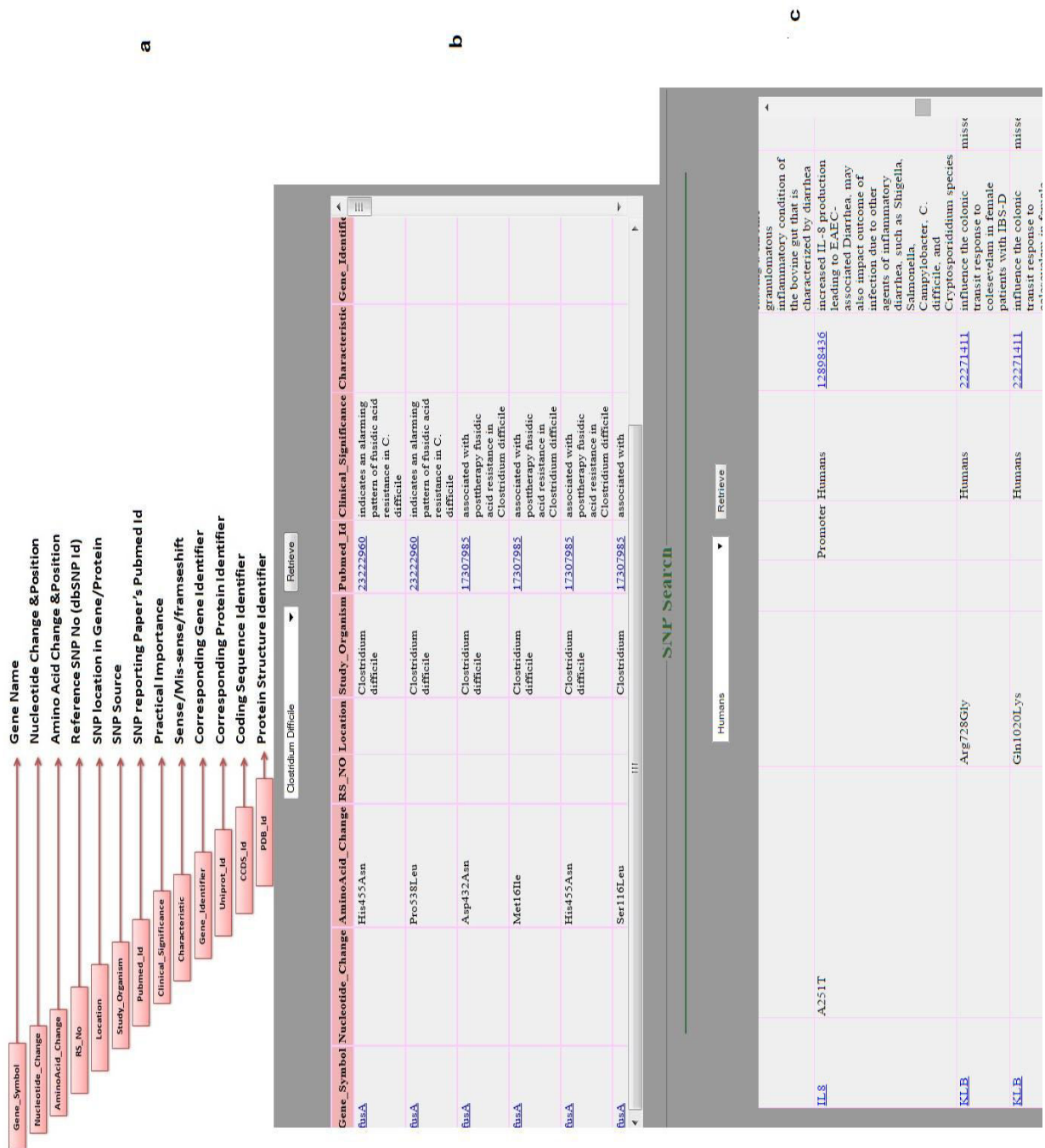


Figure 2.2: Typical output of a query (SNP). (a) Description of different annotation terms; (b) Distribution of mutations associated with antibiotic resistance in *C. difficile*; (c) Mutation in IL8 associated with susceptibility to EAEC-associated traveler's diarrhea in *H. sapiens*.

Resistance Genes Search								
<div> <div>Campylobacter Spp.</div> <div>Retrieve</div> </div>								
Gene_Name	Protein_Name	Location	Organism	Geographical_Location	Resistance_Mechanism	Clinical_Significance	Pubmed_Id	Uniprot_Id
aadA	Aminoglycoside adenylyltransferase	Plasmid	Campylobacter spp.	France	Plasmid Mediated	confers resistance to streptomycin	2168151	Q8KRK6
aadE	Streptomycin aminoglycoside 6-adenylyltransferase	Plasmid	Campylobacter spp.	France	Plasmid Mediated	confers resistance to streptomycin and spectinomycin	2168151	Q4VRA5
ant(6)-Ib	Streptomycin aminoglycoside adenylyltransferase ant(6)-Ib	Plasmid	Campylobacter fetus subsp. Fetus	Switzerland	Plasmid Mediated	confers resistance to streptomycin	20479200	D8L9Z0
cmeB	RND efflux system, inner membrane transporter CmeB	Efflux Pump (cmeABC)	Campylobacter spp.	United States	Efflux pump	resistance to ciprofloxacin, erythromycin, and tetracycline	16048946	Q5HWA1
cmeC	Outer membrane channel protein CmeC	Efflux Pump (cmeABC)	Campylobacter jejuni	United States	Efflux pump	confers resistance to cefotaxime and erythromycin	15728904	Q0PBE5
cmeR	Transcriptional repressor	Efflux Pump (cmeABC)	Campylobacter jejuni	United States	Efflux pump	confers resistance to cefotaxime and erythromycin	15728904	Q7B8P6
dfi1	Dihydrofolate reductase	Integron 1	Campylobacter jejuni	Sweden	Integrans and Gene Cassettes	confers resistance to trimethoprim	10990262	Q9EXQ6
OXA-61	OXA-61 beta-	Plasmid	Campylobacter	Australia	Plasmid Mediated	confers resistance to	15917560	

Figure 2.3: Typical output of a query (Resistance Gene). This view shows the distribution of resistance genes distributed in *Campylobacter* spp.

Users can also browse through the resistance genes in different diarrheal pathogens. Figure 2.3 shows the results of a query executed with *Campylobacter* spp. listing all the resistance genes and related annotation data. Users can also search for the related literature based on categories polymorphism, plasmid-mediated resistance or integrons and gene-cassettes. In addition to presenting the literature in a manually curated database, this web interface will also serve as a gateway to post data from future research. In keeping with the scientific philosophy of open access, the database is publicly accessible. We hope that this database will serve as a portal to share future research on antibiotic resistance.

2.4 Future directions

The current version of the database is based on literature available till date. The database will be updated regularly to provide an accurate picture of the mutations and resistance genes associated with antibiotic resistance in diarrheal pathogens. To keep the database updated, we will include

new mutation prevalence data sets. We will include novel mutations from published studies in clinical isolates of drug-resistant diarrheal pathogens. We will customize the website to house more plasmid mediated resistance and integrons and gene cassettes associated resistance mechanisms and related mutations in near future.

Conclusion

In this manuscript we describe DBDiaSNP, a comprehensive database of molecular variations and resistance genes in clinical isolates of Diarrhea, generated through manual curation of data from published literature. DBDiaSNP is intended to pace up with research efforts for global surveillance and control of diarrhea. The variations reported in the database represent potential markers of antibiotic resistance in diarrheal pathogens which might help in understanding genotype-phenotype relationships. The database will expand as additional polymorphisms will be identified in the coming years. We expect that it will make a substantial contribution towards the breakthroughs that will find applications ranging from diagnostics to drug discovery and will prove to be utile to both the diarrheal research community and clinicians. For future translational research involving integrative biology and global health [26,27], the presented database offers veritable potentials, particularly for developing countries and worldwide monitoring and personalized effective treatment of pathogens associated with diarrhea.

References

1. Atia AN, Buchman AL. Oral Rehydration Solutions in Non-Cholera Diarrhea: A Review. *Am J Gastroenterol*, 104(10), 2596-2604 (2009).
2. Harbottle H, Thakur S, Zhao S, White DG. Genetics of Antimicrobial Resistance. *Animal Biotechnology*, 17(2), 111-124 (2006).
3. Alekshun MN, Levy SB. Molecular Mechanisms of Antibacterial Multidrug Resistance. *Cell*, 128(6), 1037-1050 (2007).
4. Sobrino B, Bri0n M, Carracedo A. SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic science international*, 154(2-3), 181-194 (2005).
5. Cai M, Dai H, Qiu Y *et al.* SNP set association analysis for genome-wide association studies. *PloS one*, 8(5), e62495-e62495 (2013).
6. Yao C, Leng N, Weigel KA, Lee KE, Engelman CD, Meyers KJ. Prediction of genetic contributions to complex traits using whole genome sequencing data. *BMC Proceedings*, 8(Suppl 1), S68-S68 (2014).

7. Pinto N, Gamazon ER, Antao N *et al.* Integrating cell-based and clinical genome-wide studies to identify genetic variants contributing to treatment failure in neuroblastoma patients. *Clinical pharmacology and therapeutics*, 95(6), 644-652 (2014).
8. Chakravarti A. Single nucleotide polymorphisms: . . .to a future of genetic medicine. *Nature*, 409(6822), 822-823 (2001).
9. Bennett PM. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *British Journal of Pharmacology*, 153(S1), S347--S357 (2008).
10. Baux D, Faugere V, Larrieu L *et al.* UMD-USHbases: a comprehensive set of databases to record and analyse pathogenic mutations and unclassified variants in seven Usher syndrome causing genes. *Human mutation*, 29(8), E76-87 (2008).
11. Bharti R, Das R, Sharma P, Katoch K, Bhattacharya A. MTCID: a database of genetic polymorphisms in clinical isolates of Mycobacterium tuberculosis. *Tuberculosis (Edinburgh, Scotland)*, 92(2), 166-172 (2012).
12. Chernyaeva EN, Shulgina MV, Rotkevich MS *et al.* Genome-wide Mycobacterium tuberculosis variation (GMTV) database: a new tool for integrating sequence variations and epidemiology. *BMC genomics*, 15(1), 308-308 (2014).
13. Dai E, Lv Y, Meng F *et al.* CREAM: a database for chemotherapy resistance-associated miRSNP. *Cell death & disease*, 5, e1272-e1272 (2014).
14. Liu B, Pop M. ARDB--Antibiotic Resistance Genes Database. *Nucleic acids research*, 37(Database issue), D443-447 (2009).
15. McArthur AG, Wagglechner N, Nizam F *et al.* The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy*, 57(7), 3348-3357 (2013).
16. Stucki D, Gagneux S. Single nucleotide polymorphisms in Mycobacterium tuberculosis and the need for a curated database. *Tuberculosis (Edinburgh, Scotland)*, 93(1), 30-39 (2013).
17. Vishnoi A, Srivastava A, Roy R, Bhattacharya A. MGDD: Mycobacterium tuberculosis genome divergence database. *BMC genomics*, 9, 373-373 (2008).
18. Ramana J, Tamanna. dbDiarrhea: the database of pathogen proteins and vaccine antigens from diarrheal pathogens. *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 12(8), 1647-1651 (2012).
19. Berman HM, Westbrook J, Feng Z *et al.* The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235-242 (2000).
20. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl 1), D514-D517 (2005).
21. Landrum MJ, Lee JM, Riley GR *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, (2014).

22. Sherry ST, Ward MH, Kholodov M *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1), 308-311 (2001).
23. Palzkill T. Mitigating Antibiotic Resistance with DNA Sequence Information. In: *Emerging and Persistent Infectious Diseases: Focus on Antimicrobial Drug Resistance*. (Institute on Science for Global Policy (ISGP) Baylor College of Medicine, Houston, Texas, 2013) 19-22.
24. Bennett DE, Camacho RJ, Otelea D *et al.* Drug Resistance Mutations for Surveillance of Transmitted HIV-1 Drug-Resistance: 2009 Update. *PLoS ONE*, 4(3), e4724 (2009).
25. Voisey J, Morris CP. SNP technologies for drug discovery: a current review. *Current drug discovery technologies*, 5(3), 230-235 (2008).
26. Ahmed AA, Mohamed AA, Guled IA, Elamin HM, Abou-Zeid AH. Knowledge Translation in Africa for 21st Century Integrative Biology: The "Know-Do Gap" in Family Planning with Contraceptive Use among Somali Women. *OMICS: A Journal of Integrative Biology*, 18(11), 696-704 (2014).
27. Dandara C, Huzair F, Borda-Rodriguez A *et al.* H3Africa and the African Life Sciences Ecosystem: Building Sustainable Innovation. *OMICS: A Journal of Integrative Biology*, 18(12), 733-739 (2014).

CHAPTER 3

**Molecular Dynamics Simulations and Residue Interaction
Network Analysis in Enterotoxigenic *E. coli* and *C. jejuni*
gyrA mutants: Repercussions on Quinolone Affinity**

3.1 Introduction

gyrA, a type II topoisomerase, is an attractive drug target for antibacterials as it is indispensable to bacterial survival and lacking in higher eukaryotes. Due to their unique mechanism of action of catalyzing change in DNA topology and indispensable nature, gyrA has been exploited as a drug target both for gram negative and some gram positive bacteria since a long time. It is one of the two essential subunit of DNA gyrase, which facilitates DNA unwinding at replication forks and involved in the regulation of DNA supercoiling [1]. DNA gyrase is a heterotetramer (A_2B_2) which is composed of two subunits of GyrA (Mol Wt.: 97 kDa) and GyrB (Mol Wt.: 90 kDa) encoded by genes *gyrA* and *gyrB* respectively [2]. A 59 kDa (GyrA59) N-terminal domain (NTD) and a 38 kDa C-terminal domain (GyrA-CTD) associate to form GyrA subunit. Currently there exist no high-resolution structure for the gyrase holoenzyme (A_2B_2) but X-ray crystal structure of the GyrA59 fragment from *E. coli* (PDB ID: 1ab4) was solved with a resolution of 2.8 Å which consists of residues 30–522 [3]. The *gyrA* N-terminal region (NTD) which cleaves single-stranded DNA has three distinctive domains (Figure 3.1). The first domain contains a helix-turn-helix (HTH) motif similar to that of the *E. coli* catabolite activator protein (CAP). It contains both active-site tyrosine which is responsible for breakage and religation of the DNA [3,4] and quinolone resistance determining region (QRDR). The second domain is termed as the “tower” and adopts an extended bi-lobed α/β structure. The third domain termed as dimerization domain is a dense bundle of small α -helices connected to the tower domain and the C-terminal domain (CTD) through long α -helices [3]. *gyrA* is exploited as a target of antibacterial agents with a high therapeutic potential due to its ability to introduce negative supercoils in DNA in an ATP-dependent reaction.

Quinolones, a class of broad spectrum antibacterial agents, have been used as most potent inhibitors of bacterial topoisomerases hitherto [5]. Nalidixic acid, the prototypical compound of first generation shows relatively narrow antibacterial spectrum and low bioavailability. Introduction of fluoroquinolones by combining the fluoro-substituent at C6 and the saturated nitrogen-containing heterocycle at C7 was a major breakthrough. Norfloxacin and its derivatives ciprofloxacin and ofloxacin were representatives of this generation with broad spectrum activity and improved pharmacokinetic profiles [5]. Quinolones exert their antibiotic effect by stabilizing the covalent enzyme–DNA complex and preventing the religation of the cleaved DNA and

forming DNA lesions [6] which leads to impaired fundamental cellular activities such as DNA replication, recombination, transcription, and thus hampers bacterial growth.

Traveler's Diarrhea (TD) is the most frequent health problem affecting visitors to developing countries and tropical areas with huge economic costs. Enterotoxigenic *E. coli* (ETEC) is a major cause of TD [7]. Fluoroquinolones are potent antimicrobial agents for treatment of severe cases of TD. Ciprofloxacin is one of the most widely used fluoroquinolone. Increased use of quinolones led to an escalating rate of antibiotic resistance against these antibiotics in *E. coli*. *C. jejuni* in recent decades has emerged as the most recurring cause of infectious diarrhea infecting estimated 2.4 million people worldwide and contributes to a huge economic burden [8]. Campylobacteriosis, the most common form of enteric bacterial infections, is usually self-limiting, but in persistent form require antibiotic treatment with erythromycin and fluoroquinolones. The growing body of evidences has documented resistance to different antibiotics such as tetracycline, kanamycin, chloramphenicol, erythromycin, ciprofloxacin in *C. jejuni* strains [9,10]. Various studies on clinical isolates of ETEC and *C. jejuni* have reported mutations in the chromosomal genes that encode the subunits of DNA gyrase and topoisomerases IV which leads to increased Minimum Inhibitory Concentration (MIC) values of fluoroquinolones [11-13]. This observation suggests that the drug resistance may be related to the specific interactions between the antibiotic and the target. Increased quinolone resistance is related to i) decreased influx of antibiotic ii) plasmid mediated resistance and iii) acquisition of mutation in genes encoding gyrase reckoned to be most clinically relevant [5,14]. These mutations have been mapped to the QRDR of *E. coli* gyrase subunit A (residues 67–106) [15].

Such mutations can cause structural modifications of the target and ultimately affect binding affinity. Exploring the structural impact of these antibiotic resistance associated mutations could significantly advance our knowledge of associated molecular changes in the functionally significant regions of the target. S83L and double mutation S83L/D87N have been widely associated with quinolone resistance in ETEC isolates and T86I and a double mutation T86I/P104S in *C. jejuni* isolates of diarrheal patients from diverse geographic regions and documented to increase Minimum Inhibitory Concentration (MIC) values [9,10,16,17]. Applying molecular dynamic simulation approaches to study these mutations could yield insights into the underlying molecular mechanism of the associated phenotypes. Structural consequences of these

mutations in the target can induce functional changes. Apparently conformational flexibility of a protein affects its interaction with ligand. The present study was undertaken to probe the dynamic behavior of the gyrA in ETEC and *C. jejuni* to expound the molecular changes associated with mutations S83L, S83L/D87N and T86I, T86I/P104S leading to ciprofloxacin resistance in ETEC and *C. jejuni* gyrA respectively.

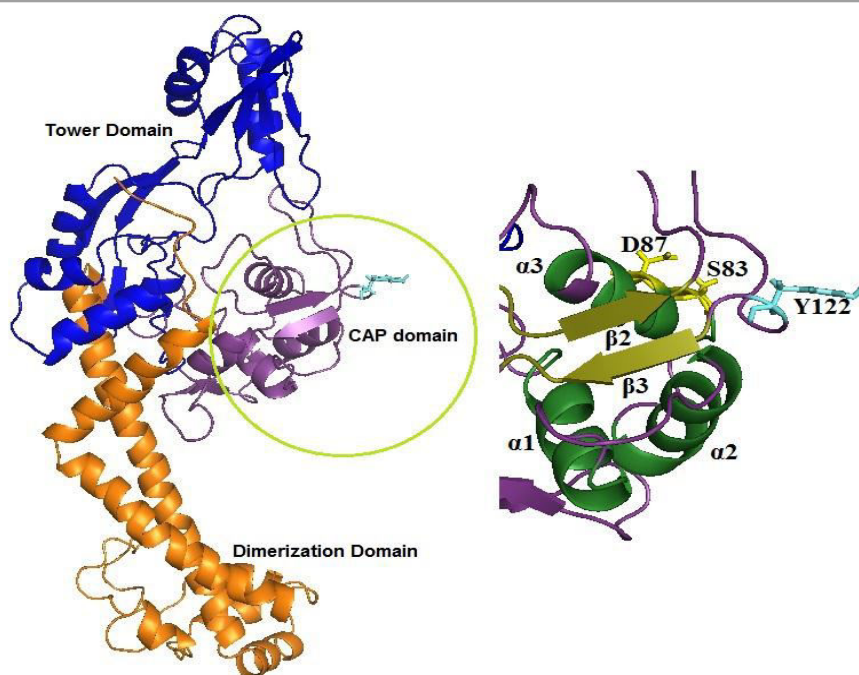


Figure 3.1: Cartoon representation of ETEC gyrA showing its distinctive domains, catalytic site and mutations associated with quinolone resistance reported in diarrheal patients. QRDR mutations and the catalytic site are in stick representation and shown in yellow and cyan color respectively. The CAP domain is colored violet purple, the tower blue, the dimerization domain and connecting α -helices colored orange. Zoomed in is a view of CAP domain showing HTH motif with three helices and two beta-sheets with corresponding residue ranges as $\alpha 1$ (43–56), $\alpha 2$ (66–77), $\alpha 3$ (81–92), $\beta 2$ (101–106), $\beta 3$ (124–128).

3.2 Methods

All the docking calculations were carried out in Accelrys Discovery Studio (DS) version 4.1 and LeadIt version 2.1.6[18]. DS Client 4.1 was used for docking preparation; LeadIt was used for

binding energy calculations. LeadIt uses FlexX algorithm which is a fast, flexible docking method that uses an incremental construction algorithm for ligand placement into active site.

3.2.1 Ligand preparation

Ciprofloxacin (Pubchem Compound ID: 2764) 2D structure (SDF format) was retrieved from Pubchem Compound database [19]. The structure was prepared by assigning appropriate ionization in physiological pH range 6.5-8.5 because at pH greater than 6.09, the carboxylic acid group will be primarily dissociated and at pH less than 8.74, the nitrogen will be primarily protonated. A conformational search for prepared ligand was performed in order to find a set of low-energy conformers. For ligand structure, a maximum of 10 tautomers and stereoisomers were generated. The prepared structure was energy minimized by Smart Minimizer Algorithm using CHARMM force field.

3.2.2 Homology modeling

gyrA amino acid sequence for ETEC (875 aa) strain E24377A (*Escherichia coli* O78:H11) and *C. jejuni* (863 aa) strain NCTC11168 were retrieved from Uniprot database with IDs A7ZP49 and Q03470 respectively. Both structures were modeled in Discovery Studio using *E. coli* gyrA (PDB Id: 1AB4) as template which showed 91.2% and 50.3% identity with ETEC and *C. jejuni* gyrA respectively. Initial homology models were built using DS. Out of the 5 models generated by DS, the one with the highest Discrete Optimized Protein Energy (DOPE) score was selected for each structure. DOPE score is a statistical measure for assessment of the quality of homology modeled structures as a whole. The residue profiles of the models were further verified for stereo-chemical properties by subjecting them to Verify3D [20] and PROCHECK [21].

The modeled native and mutant structures for ETEC and *C. jejuni* gyrA were prepared and energy minimized in DS. Mutant type structures (MT) of ETEC were obtained from wild type (WT) ETEC gyrA by introducing amino acid substitutions at position 83 and 87, generating mutants with single mutation S83L and double mutation S83L/D87N. For *C. jejuni*, two mutants were built with amino acid substitutions at 86 and 104 positions generating a single mutant T86I and a double mutant T86I/P104S. The receptor structures were prepared by optimizing hydrogen network, inserting missing loops based on SEQRES data in PDB, optimizing short and medium loop regions with looper algorithm and minimizing remaining loop regions, assigning appropriate ionization states at physiological pH 7.4. All structures were then energy minimized

to RMS gradient 0.1 with 200 steps of minimization, so as to remove steric clashes, by Smart Minimizer Algorithm with CHARMM force field. The secondary structures of the modeled gyrA from ETEC and *C. jejuni* were predicted using STRIDE [22]. The modeled structures were superimposed on the template crystal structure without changing the coordinate system of atoms in the template using jCE algorithm in Combinatorial Extension method [23].

3.2.3 Molecular docking

For ETEC and *C. jejuni*, amino acids at corresponding mutated positions 83, 87 and 86, 104 in the receptor were selected to define a sphere centre and amino acids within 10 Å radius of this sphere were selected to specify the binding site to allow the ligand to rotate freely even in its fully extended conformation. The minimized structures of ligands and receptors were used for docking in LeadIt using FlexX incremental algorithm which decomposes ligand into fragments and then selects some base fragments and use them as anchors for systematic conformational analysis of remainder of the ligand using a variety of placement strategies. For present work, we employed entropy approach which utilizes single interaction scan (SIS) based on hydrophobic pockets with only a few interaction sites. For protein ligand clashes and intra ligand clashes default values of maximum allowed overlap volume (2.9 cubic Å) and clash factor (0.6) were used. For each iteration and fragmentation, 200 solutions were generated. Top 10 best poses for the ligand were generated for each docking calculation and conformations with highest docking score (more negative) and maximum number of interactions were used for further analysis. The output score is based on Bohm's function which involves a number of parameters [24] and is calculated as:

$$\begin{aligned}\Delta G_{bind} = & \Delta G_0 + \Delta G_{hb} \sum_{h-bonds} f(\Delta R, \Delta \alpha) \\ & + \Delta G_{ionic} \sum_{ionic\ interaction} f(\Delta R, \Delta \alpha) \\ & + \Delta G_{lipo} |A_{lipo}| + \Delta G_{rot} NROT\end{aligned}$$

where ΔG values on the right side of equation are all constants, ΔG_0 is contribution to the binding energy that does not depend directly on any specific interaction with the protein, hydrogen bonding and ionic terms depending on the geometry of interaction, with deviations from ideal distance R and ideal angle α being penalized, lipophilic term $|A_{lipo}|$ proportional to the contact surface area between non polar atoms of protein and ligand, entropy term is directly

proportional to the number of rotatable bonds in ligand (NROT) and represents the penalty associated with freezing internal rotations of the ligand.

3.2.4 MD simulations

The molecular system was simulated using Molecular Dynamics in Amber 11.0 [25] and Ambertools 1.5. Hydrogen atoms were added using the tLeap program implemented in Ambertools and parameters were assigned according to AMBER FF99SB force field [26]. The system was solvated with TIP3P water model [27] in an octahedral box with 15 Å buffer around the complex. Na⁺ counterions were added to maintain the neutrality of the systems.

Each molecular system of ETEC i.e. 1) native modeled complex of 1AB4, 2) modeled complex of 1AB4 with substitution of Ser to Leu at position 83, and 3) with substitution of Ser to Leu at position 83 and substitution of Asp to Asn at position 87 and *C. jejuni* viz. 1) native modeled complex of 1AB4, 2) modeled complex of 1AB4 with substitution of Thr to Ile at position 86, and 3) with substitution of Thr to Ile at position 86 and substitution of Pro to Ser at position 104 were energy minimized through three rounds of energy minimizations of 1000 steps each. In each set of minimization, protein was energy minimized for 500 steps of steepest descent followed by 500 steps of conjugate gradient with a time step of 2 fs to eliminate close van der Waals contacts. In the first and second minimization rounds positions restraints of 10 kcal⁻¹Å⁻² and 2 kcal⁻¹Å⁻² respectively were imposed while in the third minimization, the entire system was minimized without any restraints. After minimization, each system was then gradually heated from 0 to 300 K followed by constant temperature equilibration at 300 K and 1 atm pressure for 100 ps. Following the stabilization of thermodynamic properties, a 10 ns MD simulation was performed for each system with an integration step of 2 fs in an isothermal isobaric ensemble (NPT, T = 300 K and P = 1 atm) with periodic boundary conditions. NPT conditions were maintained with Langevin thermostat [28] and Berendsen barostat [29]. SHAKE algorithm [30] was applied to constrain all covalent bonds involving hydrogen atoms.

Long-range electrostatic forces were treated using the particle-mesh Ewald (PME) method [31] with a charge grid spacing of ~1.0 Å, and the charge grid interpolation on a cubic grid by setting the direct sum tolerance to 4.0 X 10⁻⁶. Short-range electrostatics and van der Waals interactions were evaluated using a 9.0 Å atom-based cutoff. Trajectory coordinates were saved at every 2 ps for further analysis. After performing MD simulations, the root mean square deviations

(RMSD) and root mean square fluctuations (RMSF) of the MD trajectories were analyzed with respect to each initial set of coordinates. MD simulation protocol has been validated based on the modeled structure of WT-gyrA. The convergence of simulations was analyzed by monitoring parameters like the energy components, root mean-square deviation (RMSD) from the initial modeled structure, and root mean-square fluctuation (RMSF) for WT-gyrA and each of the mutants in both ETEC and *C. jejuni* molecular systems being investigated.

3.2.5 MM-PBSA calculations

Free energies were calculated using the conventional MM-PBSA and MM-GBSA approaches in AMBER 11. Free energies were estimated by averaging the configurations that were extracted from a total of 2500 frames at every 20 ps giving a total of 125 frames for energetic analysis during the last 5 ns of the MD simulation. The binding free energy profiles were calculated for WT and each of the mutants using following equations:

$$\Delta G_{\text{bind}} = G_{\text{complex}} - G_{\text{receptor}} - G_{\text{ligand}} \quad (1)$$

$$\Delta G_{\text{bind}} = E_{\text{gas}} + G_{\text{sol}} - T\Delta S \quad (2)$$

$$E_{\text{gas}} = E_{\text{int}} + E_{\text{vdw}} + E_{\text{ele}} \quad (3)$$

$$G_{\text{sol}} = G_{\text{GB}} + G_{\text{SA}} \quad (4)$$

$$G_{\text{SA}} = \gamma_{\text{SASA}} \quad (5)$$

where T and S correspond to the temperature and the total solute entropy, respectively; E_{gas} signifies gas-phase energy which is sum of internal energy (E_{int}), electrostatic (E_{ele}) and van der Waals contributions (E_{vdw}). E_{gas} is evaluated using parameters from the FF99SB force field terms. The solvation free energy (G_{sol}) can be further decomposed into polar and nonpolar solvation states. The polar solvation contribution (G_{GB} and G_{PB}) is determined by solving Poisson Boltzmann (PB) and Generalized Born (GB) equations. The nonpolar solvation contribution (G_{SA}) is estimated using $0.0072 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ as the value for constant γ and the solvent accessible surface area (SASA) determined using a water probe radius of 1.4 \AA . Dielectric constants values for solute and solvent were set to 1 and 80, respectively.

In order to determine the contribution of individual amino acid towards total binding free energy between ciprofloxacin and the WT/MT of gyrA in ETEC and *C. jejuni*, a per-residue

decomposition analysis of the interaction energy for each residue was carried out using the GB model, implemented in Amber11. Decomposition analysis was performed on 125 frames obtained during last 5 ns of MD simulation. Amino acids contributing binding free energy more than 0.5 kcal/mol were identified as the hotspot amino acids. These hotspot amino acids add up most to the complex stability and are crucial for the gyrA-ciprofloxacin interaction.

3.2.6 Principal component analysis (PCA)

In order to investigate the direction and amplitude of the dominant motions of MD trajectories, essential dynamics (ED) calculations were performed using PCA method by reducing the dimensionality of the MD simulations data. PCA can help in identifying an essential subspace that probes pronounced motions corresponding to large scale vibrational modes of groups of atoms in normal modes analysis [32]. Prior to analysis, the overall translational and rotational motions were excluded by translating MD trajectory to the geometrical center of the molecule and superimposing onto a reference structure [33]. PCA constructs the configurational space by orthogonal linear transformation in a new coordinate system to generate a covariance matrix (C). Associated eigenvectors (V_i) namely principal components (PC) and eigenvalues (λ_i) are generated by diagonalization of the covariance matrix. V_i give a vectorial description of each component of the motion by indicating the direction of the motion and λ_i represent the amplitude of the eigenvectors along the multidimensional space. Projection $\text{Proj}[M, PC_i]$ of any frame M onto i th PC is calculated by Eq. (6):

$$\text{Proj}[M, V_i] = M_{\alpha} \cdot V_i \quad 6$$

where M_{α} is the C_{α} atom of every structure after overlaying M with the reference structure. Projection highlights the time-dependent motions that the components perform in the particular vibrational mode. PCA was carried out using the PCAsuite software (<http://mmb.pcb.ub.edu/software/pcasuite/>) on I), WT, S83L, and S83L/D87N mutant models of ETEC and II) WT, T86I, T86I/P104S models of *C. jejuni*. Molecular dynamics trajectories of corresponding atoms were extracted using the PTRAJ program and analysis carried out for the last 5 ns using 63 frames.

3.2.7 Residue Interaction Networks (RINs) analysis

For both ETEC and *C. jejuni*, the average structures derived from the last 5 ns trajectory of each system were used for constructing the RINs. Reduce program [34] adds hydrogen atoms to the input structure using local geometry. PROBE [35] identifies interacting amino acids by evaluating their atomic packing using small-probe contact dot surfaces. PROBE uses several scoring functions to quantify non-covalent interactions such as interatomic contact, hydrogen bonds, salt bridges, pi-pi interaction etc. RINs generated from the averaged MD structures were visualized using RINalyzer [36]. Analyzing different RINs help us to extract changes in the networks that reflect alterations in physicochemical characteristics of the structures.

3.3 Results and discussion

ETEC

3.3.1 Model building and structure validation

As there was no crystal structure available for ETEC gyrA, it was determined using homology modeling protocol. The structure was built using *E. coli* gyrA as template which showed 91.2% identity. The sequence alignment generated using Praline program [37] is shown in Annexure: Figure 1. The modeled structure shows a typical Rossmann fold composed of six parallel beta strands linked to two pairs of alpha helices (Annexure: Figure 2). The model with lowest DOPE score was subjected to structure validation in terms of the stereo-chemical properties using PROCHECK and Verify3D. It was observed that 99.8% residues were in the allowed regions with 67.3% residues in the most favored regions (Annexure: Figure 3). Verify3D also predicted it to be a high quality model with 99.87% of the residues with score ≥ 0.2 . The structural superimposition of the backbone of the modeled structure with the template (Figure 3.2) resulted in a root mean square deviation (RMSD) of 1.3 Å with Z-score 51, indicating the structure is a reasonable good quality model. High sequence identity of the modeled structure with the template ensures correct connectivity of the secondary structure elements.

3.3.2 Molecular docking and binding affinity calculations

We used LeadIt for binding energy calculations of ciprofloxacin. FlexX incremental algorithm is aimed at assisting correct orientation of the ligands into the active site by using SIS approach. The resulting poses were ranked based on docking score. Ciprofloxacin interacting with WT gyrA through hydrogen bonds with residues K42, D87, and R91 predicted a docking score of -

25.50. Figure 3.3 captures the binding of ciprofloxacin with WT gyrA. The ability to make hydrogen bonding and hydrophobic contacts are two crucial factors that decide whether ligand fits appropriately into the binding site. After introducing mutations in the WT gyrA, docking score decreased reasonably (Table 3.1). In MT S83L and double mutant S83L/D87N docking score was observed to be -21.2327 and -18.8967 involving hydrogen bonds with K42, D87, N87 and R91 (Figure 3.4). In WT gyrA, R91 bonded with ciprofloxacin atoms O2 and O4 while in the MT gyrA hydrogen bond with ciprofloxacin O4 was lost (Table 3.2), thereby contributing to a decrease in the docking score. D87 being directly hydrogen bonded to the ciprofloxacin H43 atom, plays a crucial role in the quinolone binding with gyrA. In the double mutant hydrogen bond length for residue 87 was increased from 2.76 Å in WT gyrA to 3.25 Å significantly decreasing the bond energy.

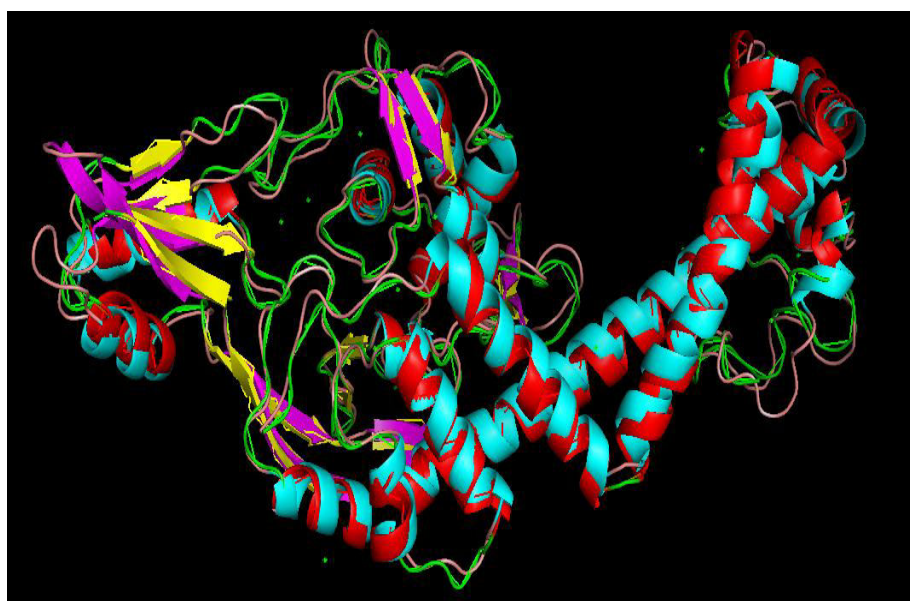


Figure 3.2: Superimposition of ETEC gyrA and the template (PDB ID: 1AB4) produced by structural modeling with color coding scheme for helix, sheet and loop as red, yellow, green in template structure and cyan, magenta, salmon in the modeled structure.

Other than hydrogen bonds, hydrophobic interactions also contribute to the docking score. Fluorine atom of ciprofloxacin made close contacts with D87. In WT gyrA complexed with ciprofloxacin, close atom-atom contacts score (clash) was 2.45, which was lower as compared to S83L and S83L/D87N in which it was observed to be 3.24 and 2.67 (Table 3.3). In S83L clash

score was highest because leucine being a bulkier residue, causes steric hindrance due to its side chain. In S83L lipophilic atom-atom contacts score was also lowest (-5.34) which was -5.61 and -6.47 in WT gyrA and double mutant respectively.

Table 3.1: Docking scores and hydrogen bond residues involved in ciprofloxacin binding with WT and MT gyrA.

Mutation	Score	No. of Hydrogen Bonds (residues involved)
WT	-25.50	3 (K42, R91, D87)
S83L	-21.23	3 (K42, R91, D87)
S83L/D87N	-18.90	3 (K42, R91, N87)

Table 3.2: Hydrogen bonds involved in gyrA-ciprofloxacin interaction in WT and MT gyrA.

Mutations	Receptor Residue	Receptor Atom	Ligand Atom	Bond Lengths (Å)
Wild Type	K42	NZ	O3	3.04
	R91	NH1	O4	2.79
	R91	NH1	O2	3.01
	D87	OD2	H43	2.76
S83L	K42	NZ	O3	2.74
	R91	NH1	O2	2.75
	D87	OD2	H43	2.80
S83L/D87N	K42	NZ	O3	2.96
	K42	NZ	O4	2.73
	R91	NH1	O2	2.75
	N87	OD1	H42	3.25

Table 3.3: Contribution of various energy terms in ciprofloxacin binding with WT and MT forms of gyrA.

Mutations	Match Score	Lipo Score	Ambig Score	Clash Score	Rot Score
Wild Type	-23.37	-5.61	-5.78	2.45	1.4
S83L	-21.88	-5.34	-4.04	3.24	1.4
S83L/D87N	-16.82	-6.47	-5.08	2.67	1.4

The lipophilic term is proportional to the contact surface area between the ligand and the protein involving non-polar terms and it contributes positively to the binding affinity. As a result binding

affinity is reduced with a decrease in the lipophilic contacts. Hydrogen bonded interactions of ciprofloxacin with both the mutants are shown in Figure 3.4.

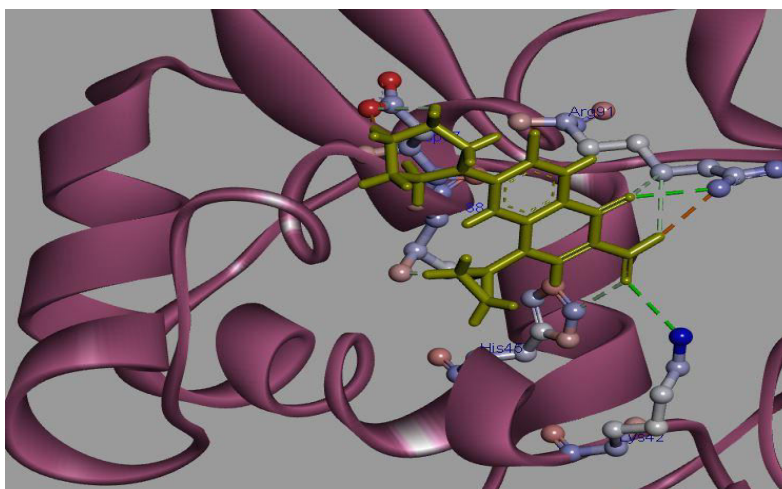


Figure 3.3: Highest scoring pose of ciprofloxacin at active site of ETEC WT gyrA. Interacting residues and atoms are labeled. Receptor is shown as cartoons. Ligand is shown in stick mode, gyrA interacting residues are shown in ball and stick mode. Binding interactions are shown as dashed lines indicating salt bridge and hydrogen bonding interactions.

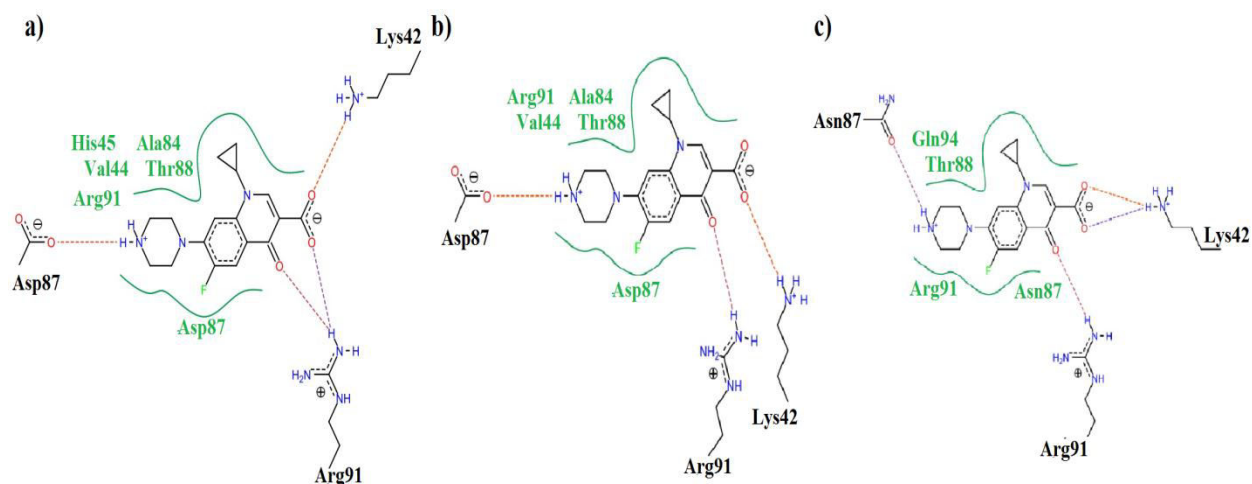


Figure 3.4: Hydrogen bonds involved in binding of ciprofloxacin with **a)** wild type **b)** mutated S83L and **c)** S83L/D87N QRDR of gyrA.

It is apparent that substitution of both S83 and D87 accounts for a decrease in the docking score and less efficient binding which ultimately leads to intermediate to high level drug resistance in

the strains previously sensitive to drugs. Our *in silico* docking results are in good agreement with previous experimental studies on these mutations [17,38,39]. Through molecular modeling studies we can derive a possible explanation for increased resistance in pathogens, but we cannot firmly establish the predictive powers of this approach since other mechanism can as well contribute to resistance such as plasmid mediated resistance, or presence of integrons and gene cassettes.

3.3.3 Structural stability characterization

To explicate the molecular basis of how the substitutions at amino acid position 83 and 87 affect susceptibility to quinolones, we carried out molecular dynamics simulations. To evaluate the conformational flexibility of the WT and MT structures, time dependent root-mean-square deviations (RMSD) of the backbone atoms from the corresponding energy minimized structure were monitored. Figure 3.5 depicts RMSDs of WT and MT models compared to the reference structure as a function of time. RMSDs were calculated from the last 5 ns of the 10-ns MD simulations. The trajectories for all the simulation systems were stabilized over the time course, with moderately small fluctuations in the models. The final RMSD values range from 1.2 to 1.3 Å for all three systems. A narrow range of time averaged RMSD values reflects that deviations in the native and mutant structures were relatively constant. It was proposed that these mutations do not disrupt the backbone conformation or cause significant global change, hence could be affecting the dynamic behavior of mutant complexes which leads to reduced binding affinity. From their initial conformations, S83L/D87N showed maximum deviation. It is evident from the RMSD plot that RMSDs of the both MT models are comparatively higher than that of the WT. In order to assess the effect of these mutation on dynamic behavior of residues, root mean square fluctuations (RMSFs) between native and S83L, and S83L/D87N were generated and plotted in Figure 3.6. RMSFs were calculated from average structures generated during the last 5 ns of the 10-ns MD simulations which were used as a reference to calculate the RMS deviations in the last 5 ns. The apex points of RMSFs were different for all the trajectories which implied that regardless of the structural similarities between WT and MT models of gyrA (RMSDs ~ 1.2 Å), flexibilities of mutant forms considerably deviate from the wild-type. The deviations in residue fluctuations may influence the interaction with ligand and thus the ligand binding efficiency. The regions of high RMSFs were not confined to vicinity of the mutated residues only, but were distributed across entire protein backbone. For example, RMSFs of the residues ranging from

240-270 show large fluctuations in WT as compared to both MT models which suggests that these mutations affect entire protein structure and not just adjacent residues. The difference in RMSDs of WT and MT models at several apex points was higher than 1 Å which further strengthen our argument that such mutations introduced in native structure influence the dynamic properties of the protein. Flexibility of active site residue Y122 is reduced in MT as compared to WT indicating the relative rigidity of this region in MT models. Y122 lies in domain III which has folds similar like *E. coli* catabolite-activator protein (CAP) and is responsible for DNA cleavage [40].

3.3.4 Secondary structure content

We show in Figure 3.7 the ensemble-averaged secondary-structure populations of each amino acid residue. Though WT and MT variants conformations appear to be dominated by turns and α -helices yet regular secondary structures (helical and β -sheet structures) in small fractions are clearly traceable. The population of the pi-helix and parallel sheet is almost negligible for all the systems investigated. The helical content is conspicuously higher than the β -sheet and turn contents in all molecular systems.

No significant changes were observed in the secondary structural content in the ensemble averaged structures of MT from WT which suggests that the backbone conformations of the WT and MT complexes do not change significantly upon mutations. DNA-dependent ATPase activity of *gyrA* is governed by the enzymatic site comprising residue Y122. To investigate the dynamic structural changes of the catalytic residue and nearby active site residues in quinolone resistance associated *gyrA* mutants, we visualized the enzymatic site using VMD [41]. It was observed that the conformation of catalytic residue Y122 fluctuates in WT, S83L, and S83L/D87N mutants during the course of simulation. Three snapshots per mutant at 0, 5 and 10 ns of the molecular dynamics trajectories were saved reflecting conformational changes of Y122 (Figure 3.8). Conformational changes of Y122 were quantified by computing the phi-psi dihedral angles across WT and MT structures (Figure 3.9). Mean dihedral angles in WT *gyrA*, S83L, S83L/D87N varied from -115.06 to -122.91 and -118.32 for phi; 4.47 to 0.73 and 5.84 for psi respectively. It is noteworthy that neither RMSD, RMSF comparisons nor secondary structure assessment provide any evidence for decreased structural stability of mutated structures as compared to WT.

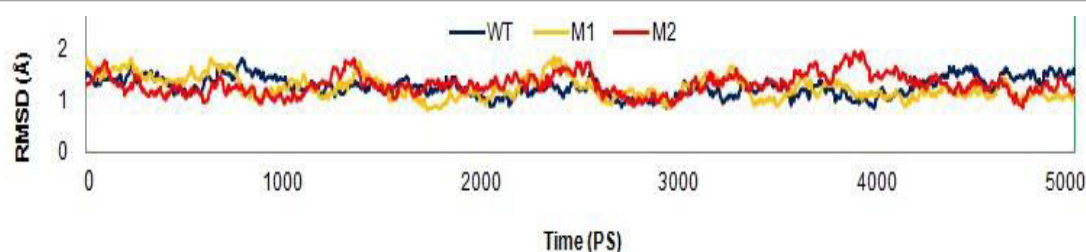


Figure 3.5: Time evolution RMSD of gyrA backbone throughout the 10 ns MD simulation time for WT (Blue), S83L mutant (Yellow:M1), S83L/D87N mutant (Red:M2).

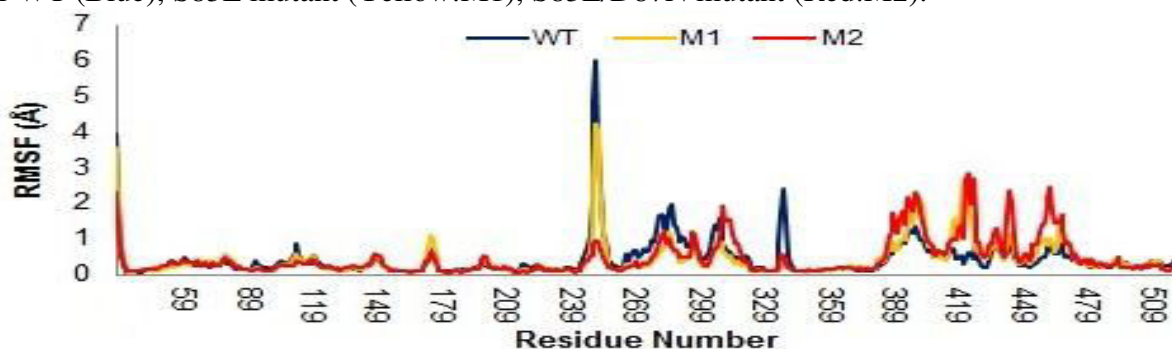


Figure 3.6: Time evolution RMSF of gyrA backbone atoms throughout the 10 ns MD simulation time for WT (Blue), S83L mutant (Yellow:M1), S83L/D87N mutant (Red:M2).

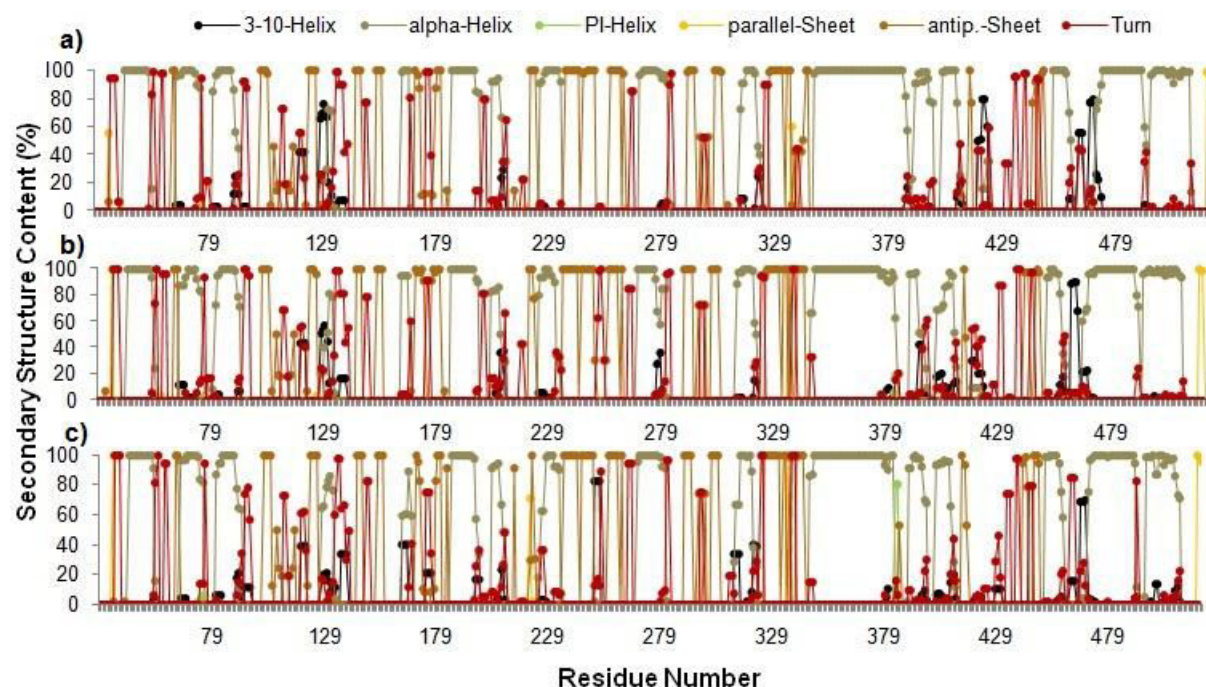


Figure 3.7: Average secondary-structure contents of protein backbone residues in a) WT, b) S83L mutant, and c) S83L/D87N double mutant.

3.3.5 MM-PB(GB)/SA binding free energy calculations

To evaluate the free energy differences between WT and MT, MM-PB(GB)/SA methods were employed. For calculating binding energy, 125 frames from the last 5 ns simulation were analyzed and predicted binding energies and energy component breakdown for different simulation systems are summarized in Table 3.4. The total binding free energies using both methods follows the same trend with a reduction in the binding affinity in MT relative to WT. In the double mutant S83L/D87N a significant drop in binding energy ($\Delta G_{\text{binding}}$) was observed using both methods. Reduced binding affinities due to the mutations influence the efficiency of ciprofloxacin binding to gyrA. Computationally predicted binding affinities are in accordance with experimental evidence that S83L and S83L/D87N lead to significant increase in Minimum Inhibitory Concentrations (MIC) values and resistance towards ciprofloxacin. The predicted van der Waals (ΔE_{VDW}) and electrostatic contribution (ΔE_{ELE}) are significantly higher in WT gyrA than both MT (Table 3.4). It is evident from Table 3.4 that majority of favorable contributions are due to ΔE_{VDW} and ΔE_{ELE} terms.

Table 3.4: MM-PB(GB)/SA based free binding energy profile of ciprofloxacin complexed with the wild and the S83L, S83L/D87N mutant types of gyrA.

Contribution	WT	S83L	S83L/D87N
ΔE_{INT}	0	0	0
ΔE_{VDW}	-24.2421	-13.4977	-5.6295
ΔE_{ELE}	-110.813	-40.6163	-33.5151
$\Delta E_{\text{GAS}} / \Delta E_{\text{MM}}$	-135.055	-54.114	-39.1446
$\Delta G_{\text{SOL-NP}}$	-2.5072	-0.8743	0.1548
ΔG_{PB}	123.253	45.0921	34.9382
$\Delta G_{\text{SOLV,PB}}$	120.746	44.2177	35.0931
$\Delta G_{\text{ELE,PB}}$	12.44	4.4758	1.4231
$H_{\text{TOT,PB}}$	-14.3093	-9.8963	-4.0515
ΔG_{GB}	121.811	46.6201	36.6038
$\Delta G_{\text{SOLV,GB}}$	118.514	44.7953	35.8509
$\Delta G_{\text{ELE GB}}$	10.9982	6.0038	3.0887
$H_{\text{TOT, GB}}$	-16.5413	-9.3187	-3.2937

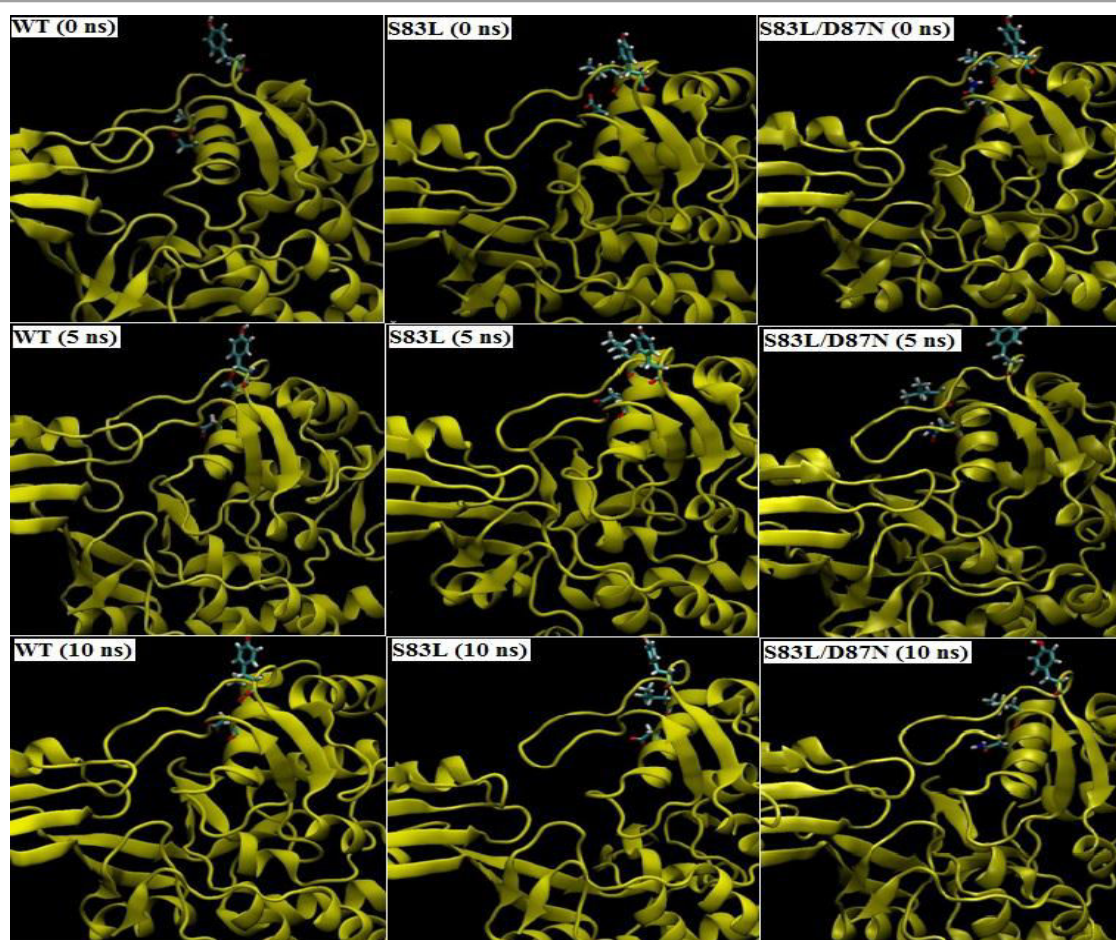


Figure 3.8: Observed conformational changes of catalytic residue Y122 and the mutated residues S83 and D87 during the course of MD simulation in WT (Column 1), S83L mutant (Column 2), S83L/D87N mutant (Column 3) at 0 ns (start of production phase), 5 ns, and 10 ns (end of production phase).

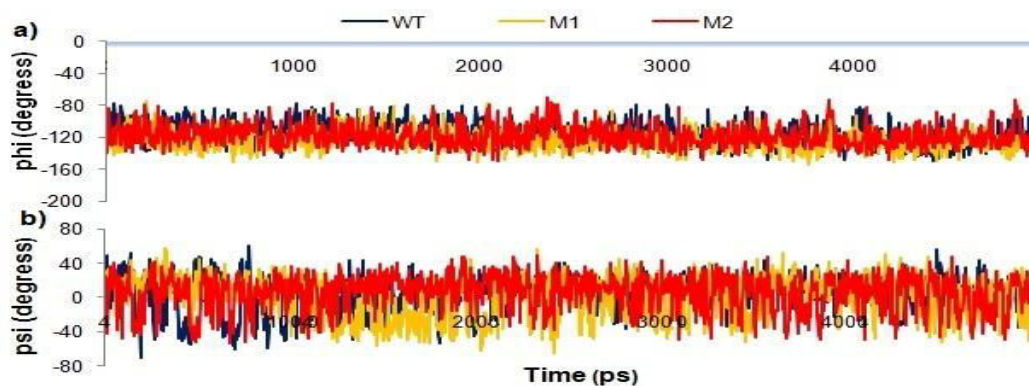


Figure 3.9: Time evolution of rotation of dihedral bond angles **a)** dihedral angle phi **b)** dihedral angle psi throughout MD simulation course in WT, S83L mutant and S83L/D87N mutant.

Relatively stable dihedral angle of catalytic residue Y122 across all structures throughout simulations suggests no major conformational change.

3.3.6 Per-residue decomposition of binding energy

Per-residue decomposition of binding energy sheds light on the residues which majorly contributes to gyrA-ciprofloxacin binding in both WT and MT models. Figure 3.10 compares the protein–ligand interaction spectra between the WT and MT models and shows contribution of individual hot-spot residues to polar and non-polar interaction terms. It can be observed from the energy decomposition analysis that in the case of ciprofloxacin bound WT complex, major contributions were from K42, H45, R91, L98, S172, G173, and I174. Decrease in the overall interaction energy in MT models is largely accountable to change in the van der Waals contribution of hot-spot residues. Although, mutation of S to L at position 83 in double mutant S83L/D87N shows improved binding at mutation site (Total Score: -0.609 kcal/mol; van der Waals Score: -0.88 kcal/mol) likely due to better hydrophobic interactions with the L side chain but it is accompanied with a reduction in overall binding energy (Table 3.4) as S83L substitution negatively impacts the binding affinity of the nearby residues.

3.3.7 PCA

To identify the dominant motions in WT and MT complexes, PCA was carried out on the last 5 ns trajectories of each molecular system. Figure 3.11 shows porcupine plot for the ciprofloxacin bound WT and MT complexes along the direction of first principal component which shows significant difference in the overall pattern of global motions between three systems. Dominant motions represented by high magnitude arrows were mostly observed in the three distinctive domains of N-terminal regions (NTD). CAP domain contains HTH motif which have been previously described to be responsible for contacts with DNA targets [40]. Motions in the CAP domain in WT were mostly upwards, while in S83L no prominent motions were observed and double mutant S83L/D87N random movements were observed. CAP domain in WT was observed to be relatively more flexible. Tower domain holds up against CAP domain providing structural support and DNA binding site [42]. Tower domain in WT exhibited mostly upward movements, while in S83L and S83L/D87N leftwards and downwards/leftwards movements were detected.

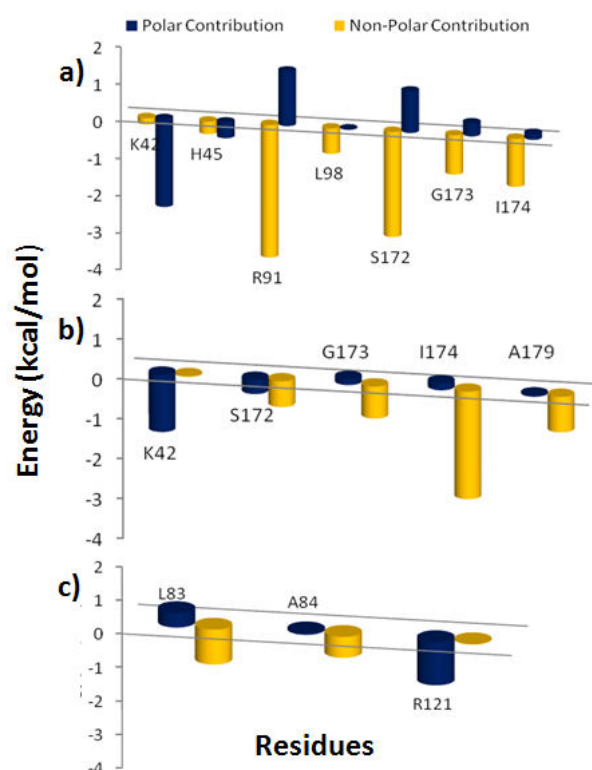


Figure 3.10: Polar and non-polar contribution of hot-spot residues from MM-PBSA based energy decomposition analysis of total binding energy (kcal/mol) during last 5 ns of MD simulation a) WT, b) S83L mutant, and c) and S83L/D87N double mutant.

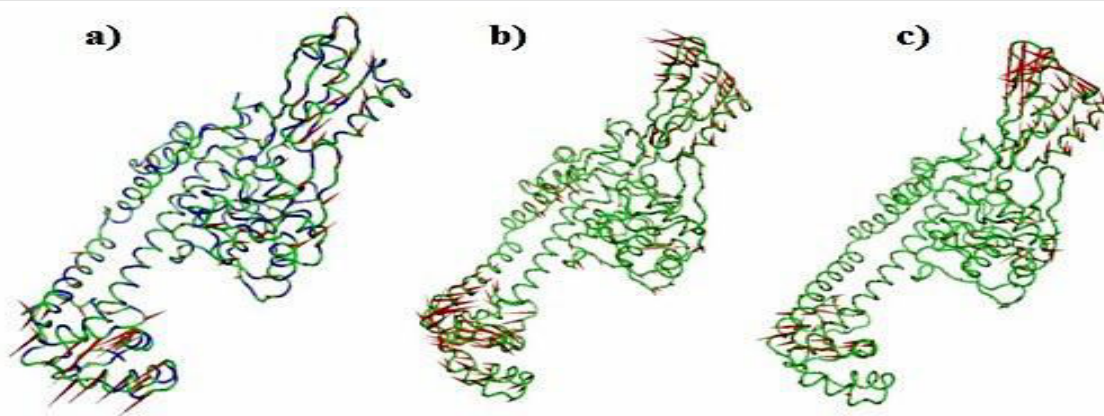


Figure 3.11: Porcupine plot of the first eigenvector obtained by PCA from simulation ensembles of a) WT, b) S83L mutant, and c) S83L/D87N mutant from the last 5 ns of MD simulation. Protein backbone represented as ribbon. Dominant motions of residues illustrated as arrows. Arrow represents direction of the eigenvector and the size of each arrow shows the magnitude of

the corresponding eigenvalue.

gyrA switches between its open and closed conformations due to the flexion of two connecting α -helices which connect dimerization domain to CTD and tower domain [42]. We observed that prominent motions in these three domains were significantly different in MT variants as compared to WT structure both in direction and magnitude. Consequently, overall conformations of trajectories from the last 5 ns simulation were different across different structures.

3.3.8 RIN analysis

To identify key residue interactions and explore the differences between different WT and MT models, RINs were generated using the representative structures from the last 5 ns trajectories of each molecular system. As evident from the RINs plot (Figure 3.12), RIN of MT variants of gyrA vary considerably when compared to the WT and each other. Zooming into the RIN plot shows for instance that interaction network in the vicinity of mutated residues S83 and D87 and around catalytic residue Y122 follows different trend across WT and MT variants of gyrA. A three way interaction (peptide bond, close atom interaction, pi-cation interaction) between T123 and E124 in WT gyrA reduces to just a peptide bond interaction in S83L mutant. Though effect of changes at individual residue site may not be prominent, but collectively these changes in residue interactions could have extreme effect on protein structure altering the active site conformation. It is quite interesting to observe that in the case of S83L/D87N mutant, interaction sub-network of residues range 30-34 collapses with Q106, T123 and E124 interaction networks while in S83L mutant, no interactions are observed in these sub-networks. S83L and D87N mutations have apparently disrupted the RIN of WT gyrA subsequently affecting the drug binding landscape.

C. jejuni

3.3.9 Model building and structure validation

Due to unavailability of crystal structure for *C. jejuni* gyrA, it was determined using homology modeling protocol. The structure was built using *E. coli* gyrA as template which showed 50.3% identity. The sequence alignment generated using Praline program is shown in Annexure: Figure 4. The secondary structure predicted using STRIDE program is shown in Annexure: Figure 5.

The model with lowest DOPE score was subjected to structure validation in terms of the stereochemical properties using PROCHECK and Verify3D. It was observed that 65.3% residues were in the most favored regions with 32.5% residues in the additional allowed regions (Annexure: Figure 6). Verify3D also predicted it to be a reasonably good quality model with 88.18% residues in agreement with 3D-1D score profile (≥ 0.2) which determines the compatibility of structure (3D) with amino acid sequence (1D). The structural superimposition of the backbone of the modeled structure with the template (Figure 3.13) resulted in a root mean square deviation (RMSD) of 0.9 Å, indicating the structure is of reasonable good quality. High sequence identity of the modeled structure with the template ensures correct connectivity of the secondary structure elements.

3.3.10 Molecular docking and binding affinity calculations

Binding energy calculations of ciprofloxacin with gyrA were carried out in LeadIt. The resulting poses were ranked based on docking score. Ciprofloxacin interacting with WT gyrA through hydrogen bonds with residues R35, and E158 predicted a docking score of -11.19. Figure 3.14 captures the binding of ciprofloxacin with WT gyrA. Hydrogen bonds and hydrophobic contacts are factors critical to accurate ligand binding into the active site moiety. Due to mutations introduced, docking score decreased reasonably (Table 3.5). In MT T86I and double mutant T86I/P104S docking score was observed to be -10.91 and -10.88 involving hydrogen bonds with H48, R94, and I177 (Figure 3.14).

In WT gyrA R35 bonded with ciprofloxacin atoms O3 and O4 while in the mutant forms of gyrA hydrogen bond with R35 was lost (Table 3.5), thereby contributing to a decrease in the docking score. Consequently it can be speculated that R35 being directly hydrogen bonded plays a crucial role in the quinolone binding with gyrA. In T86I, R94 atoms NE and NH2 interacts with ligand atom O3 with respective bond lengths of 2.57 Å and 2.92 Å while in the double mutant one hydrogen bond through gyrA NE atom was lost while the hydrogen bond length for atom NH2 was reduced. Hydrogen bond length for H48 which interacts through ligand O2 atom was reduced to 3.32 in T86I/P104S from 2.85 in T86I. The reduced bond lengths for H48 and R94 resulted in decreased binding affinity scores.

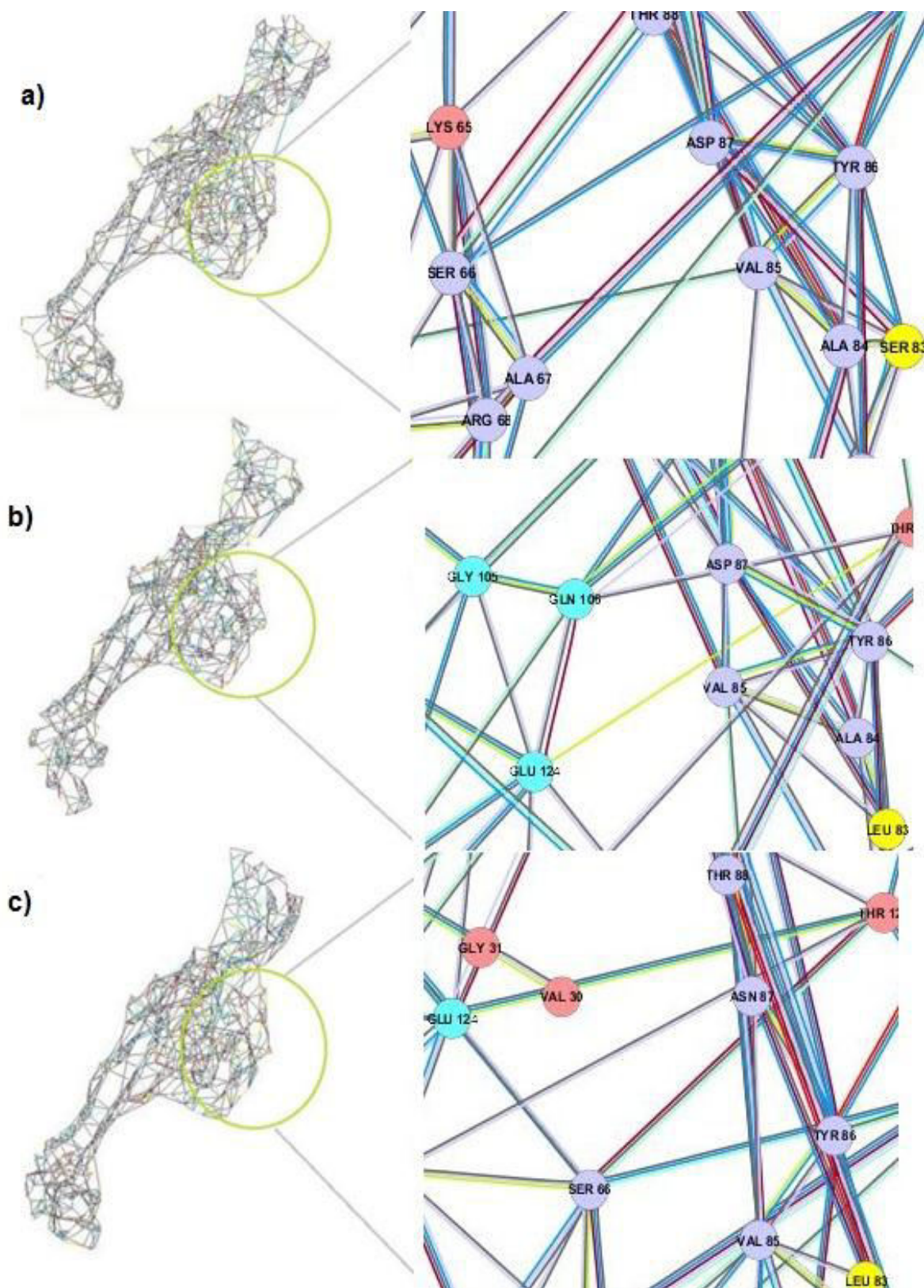


Figure 3.12: Residue interaction networks (RINs) comparison between **a)** WT, **b)** S83L mutant, and **c)** S83L/D87N mutant docked with ciprofloxacin. Zoomed view highlights changes in network in vicinity of catalytic residue Y122 and mutated QRDR residues S83 and D87.

Nodes represented as: Helix (light blue circle) Loop (purple circle) Sheet (green circle) Default (cyan circle) and Edges represented as: Hydrogen Bond (dark red line) Peptide Bond (yellow line) Salt Bridge (red line) Ionic bond (pink line) Pi Stacking (blue line) Pi-Cation (light green line) Closest Atoms Interaction (light purple line)

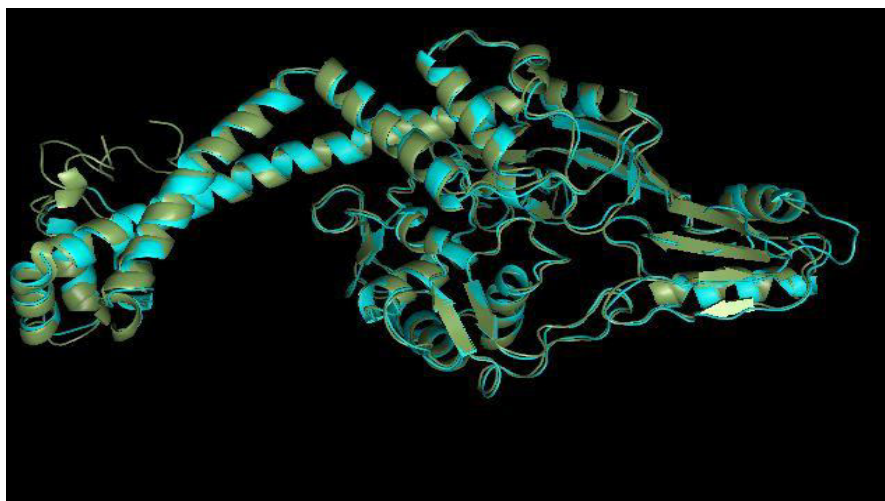


Figure 3.13 Structural superimposition of *C. jejuni* gyrA on the template (PDB ID: 1AB4) produced by structural modeling with modeled structure colored cyan and template colored sea green.

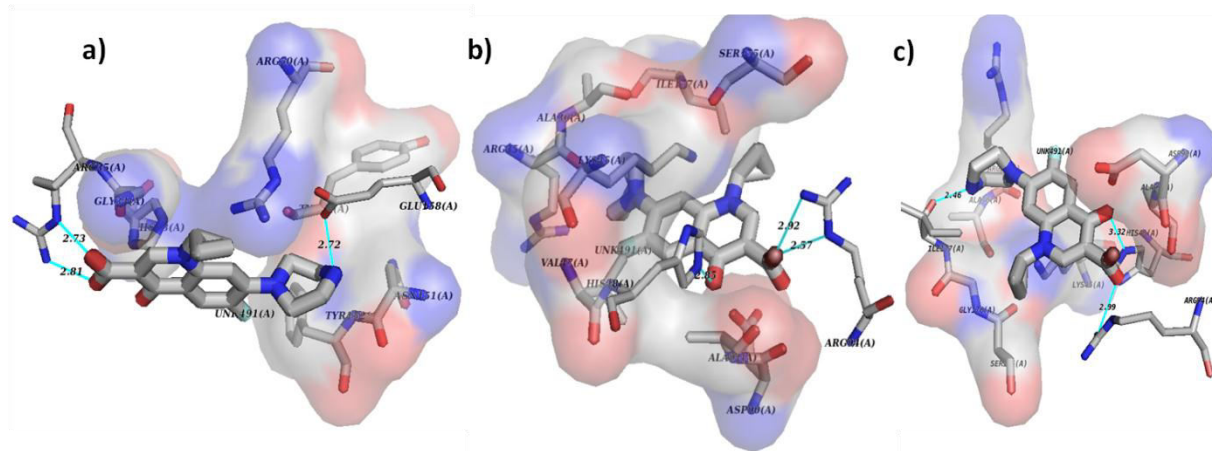


Table 3.5: Docking scores and hydrogen bond residues involved in ciprofloxacin binding with WT and mutant forms of gyrA.

Mutation	LeadIT Score (kcal/mol)	GyrA's Residue	GyrA's Atom	Ligand Atom	H-Bond Distance(Å)
Wild	-11.19	R35	NE	O3	2.73
		R35	NH2	O4	2.81
		E158	OE2	N7	2.72
T86I	-10.91	R94	NE	O3	2.57
		R94	NH2	O3	2.92
		H48	NE2	O2	2.85
T86I/P104S	-10.88	I177	O	N7	2.46
		R94	NH2	O3	2.99
		H48	NE2	O2	3.32

We can discern from these findings that substitutions of T with I at position 86 and P with S at position 104 in *C. jejuni* gyrA have a pronounced effect on ciprofloxacin binding and have been reported frequently. It is evident that mutation of both T86 and P104 resulted in a relatively low docking score which suggests the critical role of these residues in ciprofloxacin binding with gyrA. This decreasing trend in docking score accounts for increased resistance in the MT gyrA, observed to be highest in the double mutant.

3.3.11 Structural stability characterization

We carried out molecular dynamics simulations of complexes of WT and MT gyrA from *C. jejuni* with ciprofloxacin, in order to better understand the mechanism underlying gyrA resistance to fluoroquinolones at atomic level. WT and MT gyrA (T86I and T86I/P104S) docked against ciprofloxacin were computationally simulated to complement experimental findings of increased MICs reported in clinical isolates harboring these mutations. To evaluate the conformational flexibility of the WT and MT structures, time dependent root-mean-square deviations (RMSD) of the backbone atoms relative to the corresponding energy minimized structure were monitored. RMSDs of WT and MT complexes estimated relative to the reference structure are plotted in Figure 3.15 as a function of time. RMSDs were calculated from the last 5 ns of the 10 ns MD simulations. It was observed that all the molecular systems stabilized over the 10 ns time course of simulation, with moderately small fluctuations in the models arising at the beginning of every production run. For all the three systems, the final RMSD values relative

to the reference energy minimized structure range from 1.3 to 1.8 Å for all three systems. RMSD deviations in the native and mutant forms of gyrA were relatively constant which is mirrored by a narrow range of time averaged RMSD values. It suggests that these mutations do not disrupt the backbone conformation or cause significant global change. It is speculated that due to these mutations, changes in the dynamic behavior of gyrA-ciprofloxacin complexes leads to reduced binding affinity.

Root mean square fluctuations (RMSFs) of each residue in WT and T86I, and T86I/P104S mutants of gyrA were plotted in Figure 3.16 to monitor the effect of these mutations on dynamic behavior of complexes. RMSFs were computed relative to the average structures generated during the last 5 ns of the 10 ns MD simulations. The apex points of RMSFs were of different magnitude for WT and both MT trajectories which imply that despite the fact that no considerable distortion was observed in the backbone structure, residue flexibilities in mutant forms show large deviations from those observed in WT. The deviations in residue fluctuations may influence the interaction with ligand and thus the ligand binding efficiency. The regions of high RMSFs were not confined to vicinity of the mutated residues only, but were localized and distributed across entire protein backbone which suggests that effect of mutations is not restricted to its close vicinity but affect the entire protein structure. At some points, RMSF difference between WT and MTs was significantly higher which further strengthen our argument that mutations introduced in the native structure impact the dynamic behavior of the protein. Residue flexibility was observed to be lowest in T86I/P104S which would ultimately lead to reduced binding affinity.

3.3.12 Secondary structure content

The ensemble-averaged secondary-structure populations of each amino acid residue for all three molecular systems are plotted in Figure 3.17. Though WT and MT variants conformations appear to be dominated by α -helices yet turns and anti-parallel β -sheet structures were apparently visible. Other regular secondary structures were present in very small fractions. Pi-helix population was almost negligible for all the systems investigated and was almost absent from T86I mutant. The helical content was radically higher than the β -sheet and turn contents in all molecular systems.

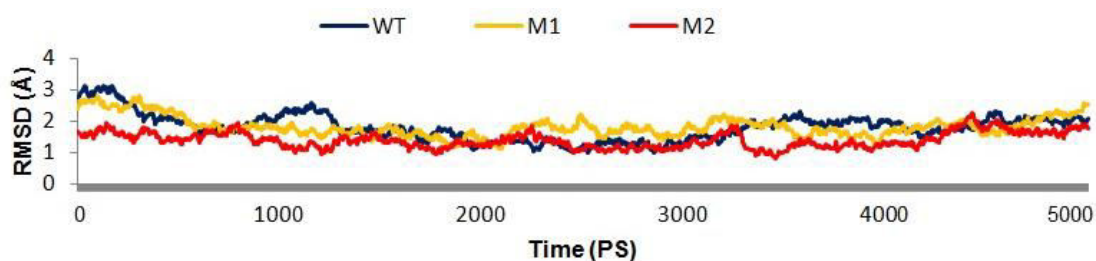


Figure 3.15: Time evolution RMSD of gyrA backbone throughout the 10 ns MD simulation time for WT (Blue), T86I mutant (Yellow: M1), T86I/P104S mutant (Red: M2).

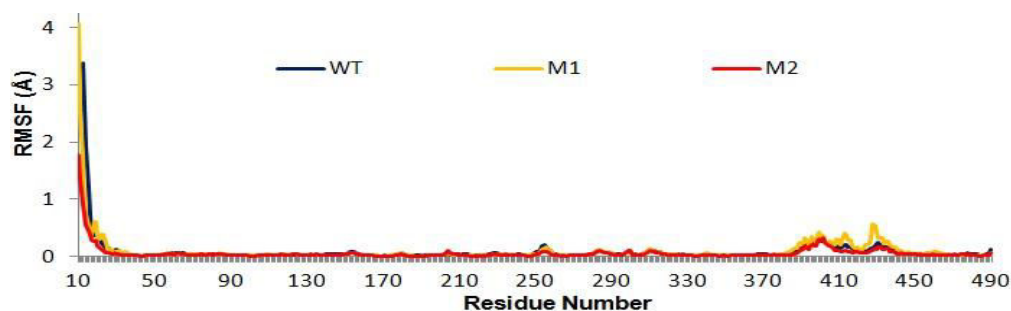


Figure 3.16: Time evolution RMSF of gyrA backbone atoms throughout the 10 ns MD simulation time for WT (Blue), T86I mutant (Yellow: M1), T86I/P104S mutant (Red: M2).

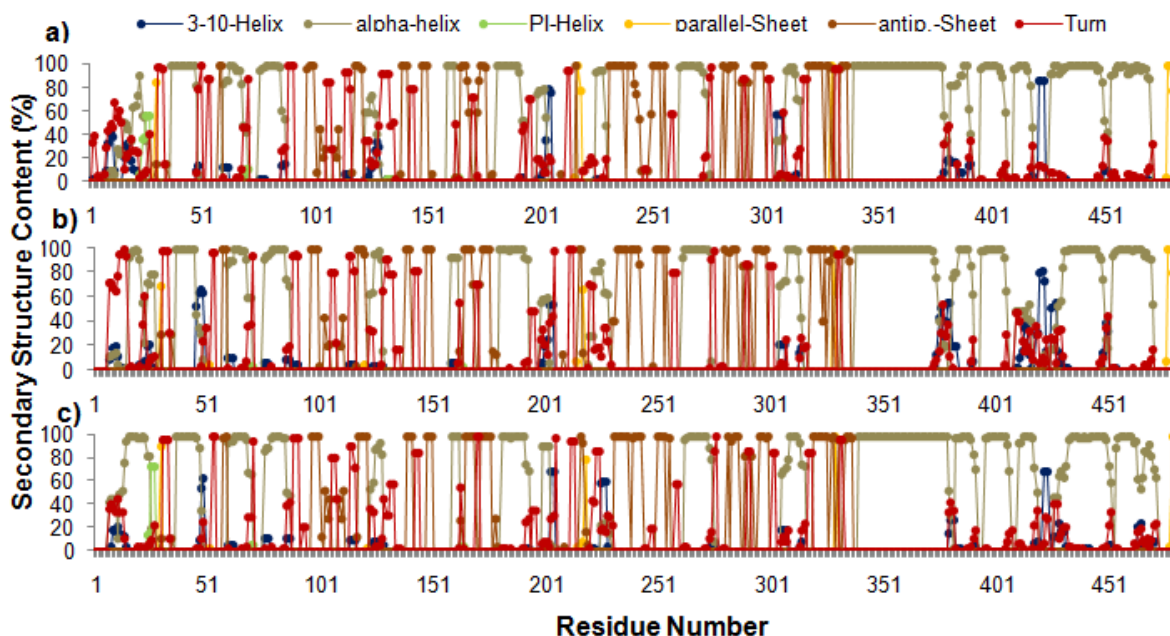


Figure 3.17: Average secondary-structure contents of protein backbone residues in a) WT, b) T86I mutant, and c) T86I/P104S double mutant.

Significant changes were not observed in the secondary structural content in ensemble averaged structures of MTs from WT which again points to the argument that the backbone conformations were not significantly altered due to mutations harbored in these structures. To investigate the dynamic structural changes of the mutated residues and overall conformation of molecular complexes, we visualized the MD trajectories using VMD [41]. It was observed that the conformation of mutated residues as well as overall backbone conformation fluctuates in WT, T86I, and T86I/P104S mutants during the course of simulation. In Figure 3.18, three snapshots saved for each molecular system at 0, 5 and 10 ns of the molecular dynamics trajectories reflecting conformational changes are shown.

3.3.13 MM-PB(GB)/SA binding free energy calculations

To evaluate the free energy differences between WT and MT, MM-PB(GB)/SA methods were employed. We estimated binding energy from 125 frames of the last 5 ns simulation trajectories and computationally predicted free energies. The contribution of various components to binding energy for different simulation systems are summarized in Table 3.6. The total binding free estimates are approximate as it does not include contribution from configurational entropy changes. In support of experimental findings, it was observed that WT gyrA has higher affinity for ciprofloxacin and binds more favorably than any of the mutant forms (Table 3.6). In both the mutants T86I and T86I/P104S a significant drop in binding energy ($\Delta G_{\text{binding}}$) as compared to WT gyrA was observed using both PB and GB methods. Reduced binding affinities due to the mutations influence the efficiency of ciprofloxacin binding to gyrA. It was revealed that electrostatic term (ΔE_{ELE}) contributes significantly to total binding free energy. Our results reinforce the idea that these mutations in the QRDR region are major driving force of fluoroquinolone resistance as computationally predicted binding affinities are in accordance with experimental evidence that T86I and T86I/P104S lead to significant increase in MIC values of various quinolones [13].

3.3.14 Per-residue decomposition of binding energy

Per-residue binding energy decomposition sheds light on the key residues which play major role in protein-ligand interaction in a molecular system and provides a semi-quantitative estimate of their contribution to complex stability. We have compared the protein–ligand interaction spectra between different molecular systems of gyrA (WT and MT models) and contribution of

individual residues classified as hot-spot, to both polar and non-polar interaction terms is plotted in Figure 3.19. From binding energy decomposition analysis, it was revealed that residues S30, R35, P82, H83, and Y152 contribute significantly to ciprofloxacin binding with WT gyrA.

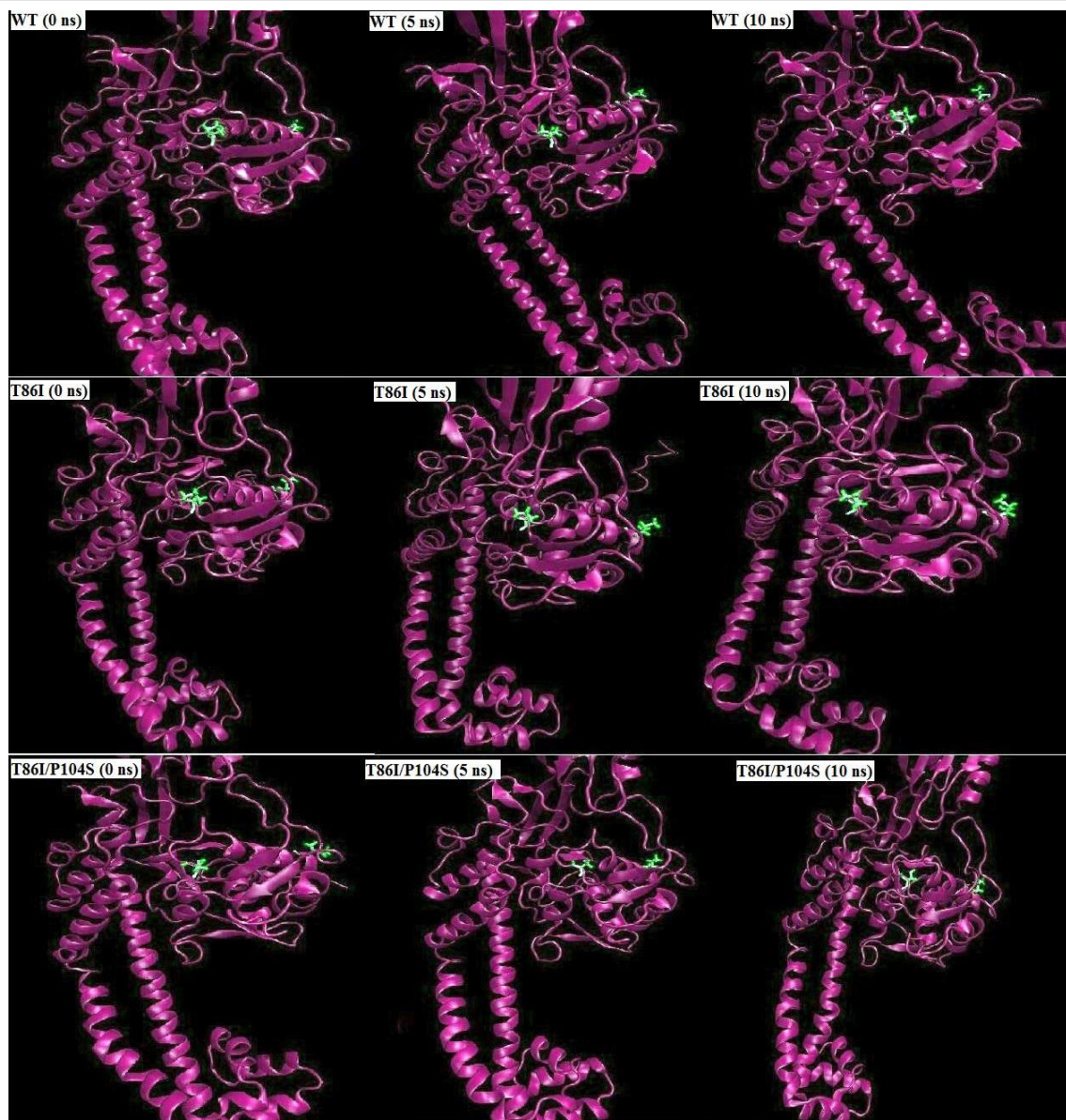


Figure 3.18: Observed conformational changes of the protein backbone and mutated residue positions 86 and 104 during the course of MD simulation in WT (Row 1), T86I mutant (Row 2), T86I/P104S mutant (Row 3) at 0 ns (start of production phase), 5 ns, and 10 ns (end of production phase).

Table 3.6: MM-PB(GB)/SA based free binding energy profile of ciprofloxacin complexed with the wild and the S83L, S83L/D87N mutant types of gyrA.

Contribution	WT	M1	M2
ΔE_{INT}	0	0	0
ΔE_{VDW}	-15.637	-26.7396	-17.2818
ΔE_{ELE}	-144.14	-85.2518	-66.3329
$\Delta E_{\text{GAS}} / \Delta E_{\text{MM}}$	-159.78	-111.991	-83.6147
$\Delta G_{\text{SOL-NP}}$	-1.3818	-2.3611	-1.6393
ΔG_{PB}	144.2091	102.7148	73.208
$\Delta G_{\text{SOLV,PB}}$	142.8272	100.3537	71.5687
$\Delta G_{\text{ELE,PB}}$	0.0685	17.463	6.8751
$H_{\text{TOT,PB}}$	-16.95	-11.6377	-12.046
ΔG_{GB}	146.058	100.406	74.49
$\Delta G_{\text{SOLV,GB}}$	143.749	96.9002	72.106
$\Delta G_{\text{ELE GB}}$	1.917	15.1542	8.1571
$H_{\text{TOT, GB}}$	-16.028	-15.0912	-11.5087

A significant decrease in the free energy in mutant forms was largely attributed to both polar and non-polar contribution of hot-spot residues. None of the hot-spot residues identified in WT gyrA were reported to have significant contribution to complex stability in mutants T86I and T86I/P104S. K45 residue in T86I/P104S has higher polar contribution as compared to T86I mutant, but R94 and S175 have higher contribution from van der Waals term in T86I. Though T86I/P104S showed improved binding at K45 site (Total Score: -4.357 kcal/mol; electrostatic Score: -33.591 kcal/mol) likely due to its hydrogen bonded interactions with the R49 side chain but it has a reduced total free energy as compared to T86I mutant (Table 3.6) as T86I and P104S substitution negatively impacts the binding affinity of the nearby residues.

3.3.15 PCA

To probe the dominant motions of ensemble of conformations in WT and MT complexes, PCA was carried out on the last 5 ns trajectories of each molecular system. The trajectories of ciprofloxacin bound WT and MT complexes analyzed using PCA to identify dominant motions along the direction of first principal component are plotted in Figure 3.20 in the form of porcupine plot. The porcupine plot represents major fluctuations and significant difference in the overall pattern of global motions in WT and MTs molecular systems are clearly distinguishable.

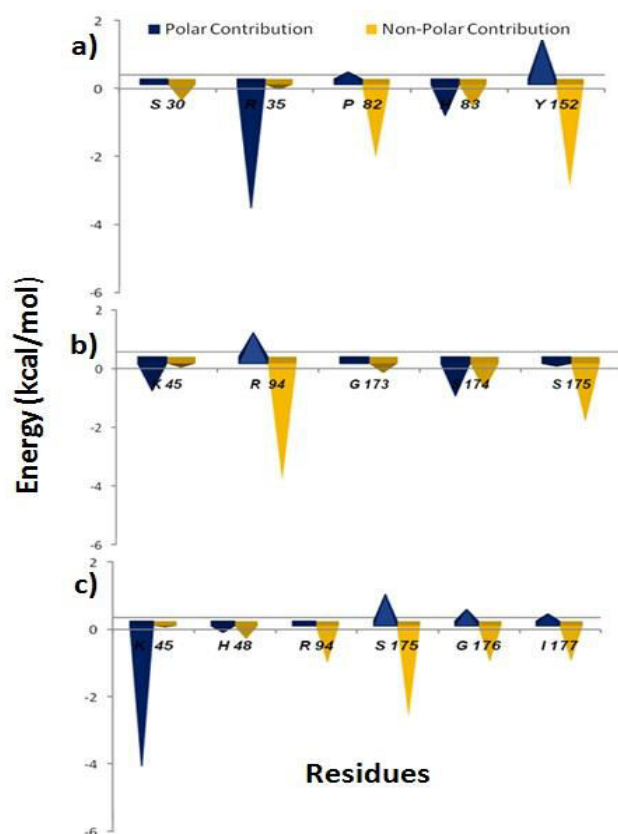


Figure 3.19: Polar and non-polar contribution of hot-spot residues in a) WT, b) T86I mutant, and c) and T86I/P104S double mutant during last 5 ns of MD simulation.

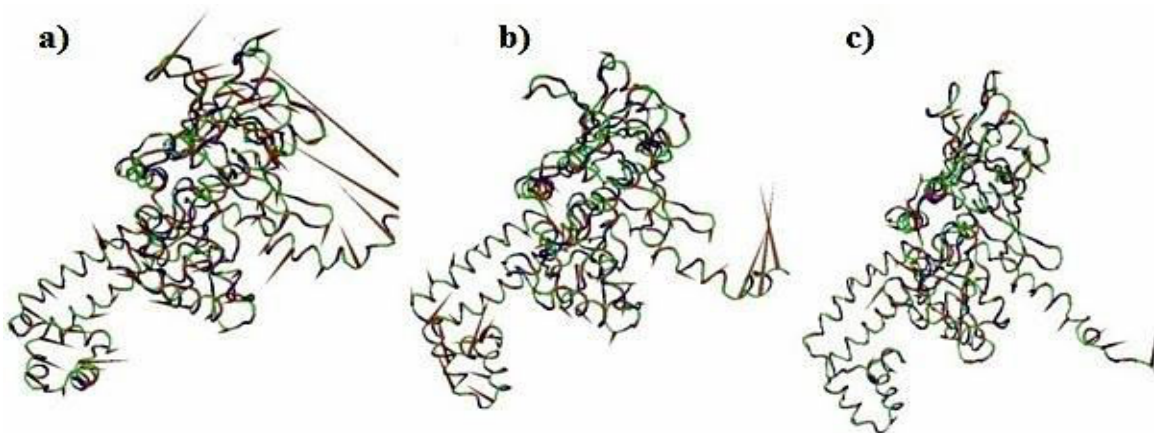


Figure 3.20: Porcupine plot of the first eigenvector obtained by PCA from simulation ensembles of a) WT, b) T86I mutant, and c) and T86I/P104S mutant from the last 5 ns of MD simulation. Protein backbone represented as ribbon. Dominant motions of residues illustrated as arrows. Arrow represents direction of the eigenvector and the size of each arrow shows the magnitude of

the corresponding eigen value.

In the WT, major dominant motions which are represented by high magnitude arrows were mostly observed in the lower and uppermost domains which are similar to tower domain in *E. coli*. These motions were of reduced magnitude and varying directions in both mutants. In T86I/P104S mutant, corresponding principal motions were observed to be of the least magnitude. Motions in the lower domain which connects two α -helices were also distinct among WT and MT forms of gyrA; T86I/P104S was characterized by prominently reduced motions which reflect reduced flexibility. gyrA is predicted to switch between its open and closed conformations due to the flexion of two connecting α -helices which connect dimerization domain to C Terminal Domain and tower domain [42]. Consequently, any changes in the flexibility of this domain would ultimately affect ligand binding efficiency. By and large, prominent dominant motions vary both in direction and magnitude among WT, T86I, and T86I/P104S mutants.

3.3.16 RIN Analysis

From both structural and functional perspectives, proteins are complex molecules and their small world behavior can be extensively studied by representing inter-residue protein contacts using graph theory in the form of residue interaction networks. RINs can capture residue interactions and find potential application in many scenarios such as identification of key residues involved in protein folding and functionally different states of protein and recently in study of protein-ligand interactions [43-45]. To identify key residue interactions and explore the differences between different WT and MT models, RINs were generated using the representative structures from the last 5 ns trajectories of each molecular system and were plotted in Figure 3.21. It is apparent that RIN plots of mutant forms of gyrA vary considerably when compared to the WT and from each other as well. A zoomed in view of the RINs shows that interaction network in the neighborhood of mutated residue positions 86 and 104 are conspicuously discernible across WT and MT variants of gyrA.

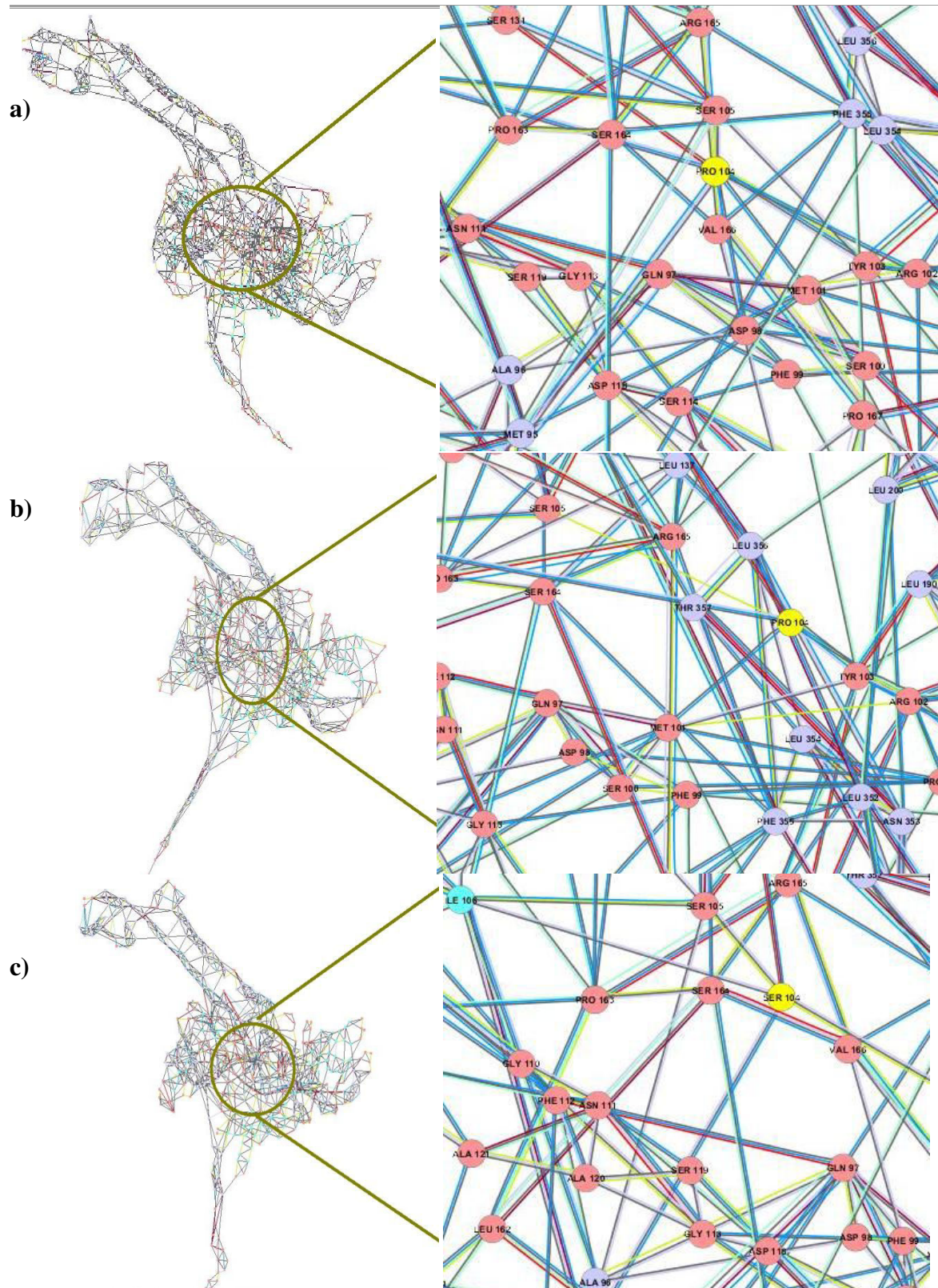








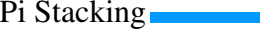
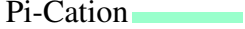
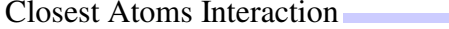


Figure 3.21: Residue interaction networks (RINs) comparison between **a)** WT, **b)** T86I mutant, and **c)** and T86I/P104S mutant docked with ciprofloxacin. Zoomed view highlights changes in network in the vicinity of mutated residue.

Nodes represented as: Helix  Loop  Sheet  Default  and Edges represented as: Hydrogen Bond  Peptide Bond  Salt Bridge  Ionic bond  Pi Stacking  Pi-Cation  Closest Atoms Interaction 

Conclusion

Rapidly evolving computational platforms have made investigating the structural consequences of the mutations easier. We performed a systematic study to explore the molecular mechanisms underlying mutations associated quinolone resistance in *gyrA* based on explicit-water MD simulations. As evidenced by RMS deviation, RMS fluctuation and secondary structure conservation, WT and MT trajectories were exploring the same conformational space and showed no significant changes. Despite that, incorporated mutations significantly impact the dynamic behavior of complexes which was reflected by major disruptions in the residue interaction network and altered dominant motions which could destabilize active-site conformation reducing ligand binding efficiency. This study offers a comprehensive picture of genotype–phenotype association of quinolone resistance associated SNPs in *gyrA*. The results reported in this study elucidate the role of mutations in *gyrA* which may provide a useful information for the design of *gyrA* mutants based therapeutic strategies against resistant strains of ETEC and *C. jejuni*.

References

1. Champoux JJ. DNA topoisomerases: Structure, Function, and Mechanism. *Annual Review of Biochemistry*, 70(1), 369-413 (2001).
2. Reece RJ, Maxwell A, Wang JC. DNA Gyrase: Structure and Function. *Critical Reviews in Biochemistry and Molecular Biology*, 26(3-4), 335-375 (1991).
3. Cabral JHM, Jackson AP, Smith CV, Shikotra N, Maxwell A, Liddington RC. Crystal structure of the breakage-reunion domain of DNA gyrase. *Nature*, 388(6645), 903-906 (1997).
4. Horowitz DS, Wang JC. Mapping the active site tyrosine of Escherichia coli DNA gyrase. *Journal of Biological Chemistry*, 262(11), 5339-5344 (1987).

5. Fabrega A, Madurga S, Giralt E, Vila J. Mechanism of action of and resistance to quinolones. *Microbial Biotechnology*, 2(1), 40-61 (2009).
6. Frederic C, Shantanu K, Anthony M. Exploiting bacterial DNA gyrase as a drug target: current state and perspectives. *Applied Microbiology and Biotechnology*, 92(3), 479-497 (2011).
7. Diemert DJ. Prevention and Self-Treatment of Traveler's Diarrhea. *Clinical Microbiology Reviews*, 19(3), 583-594 (2006).
8. Friedman CR, Neimann J, Wegener HC, Tauxe RV. Epidemiology of *Campylobacter jejuni* infections in the United States and other industrialized nations. In: *Campylobacter*. (ASM International, Washington, USA, 2000) 121-138.
9. Alfredson DA, Korolik V. Antibiotic resistance and resistance mechanisms in *Campylobacter jejuni* and *Campylobacter coli*. *FEMS Microbiology Letters*, 277(2), 123-132 (2007).
10. Thakur S, Zhao S, McDermott PF *et al.* Antimicrobial Resistance, Virulence, and Genotypic Profile Comparison of *Campylobacter jejuni* and *Campylobacter coli* Isolated from Humans and Retail Meats. *Foodborne Pathogens and Disease*, 7(7), 835-844 (2010).
11. Beckmann L, Müller M, Lubert P, Schrader C, Bartelt E, Klein G. Analysis of *gyrA* mutations in quinolone-resistant and -susceptible *Campylobacter jejuni* isolates from retail poultry and human clinical isolates by non-radioactive single-strand conformation polymorphism analysis and DNA sequencing. *Journal of Applied Microbiology*, 96(5), 1040-1047 (2004).
12. Frasao BDS, Medeiros V, Barbosa AV *et al.* Detection of fluoroquinolone resistance by mutation in *gyrA* gene of *Campylobacter* spp. isolates from broiler and laying (*Gallus gallus domesticus*) hens, from Rio de Janeiro State, Brazil. *Cienc.Rural*, 45, 2013-2018 (2015).
13. Zirnstein G, Li Y, Swaminathan B, Angulo F. Ciprofloxacin Resistance in *Campylobacter jejuni* Isolates: Detection of *gyrA* Resistance Mutations by Mismatch Amplification Mutation Assay PCR and DNA Sequence Analysis. *Journal of Clinical Microbiology*, 37(10), 3276-3280 (1999).
14. Jacoby GA. Mechanisms of Resistance to Quinolones. *Clinical Infectious Diseases*, 41(Supplement 2), S120-S126 (2005).
15. Yoshida H, Bogaki M, Nakamura M, Nakamura S. Quinolone resistance-determining region in the DNA gyrase *gyrA* gene of *Escherichia coli*. *Antimicrobial Agents and Chemotherapy*, 34(6), 1271-1272 (1990).
16. Daniel WI, Charles WH, Apichai S *et al.* Comparative Antibiotic Resistance of Diarrheal Pathogens from Vietnam and Thailand, 1996-1999. *Emerging Infectious Disease journal*, 8(2), 175 (2002).
17. Mendez AE, Pitart C, Ruiz J, Marco F, Gascon J, Vila J. Evolution of antimicrobial resistance in enteroaggregative *Escherichia coli* and enterotoxigenic *Escherichia coli* causing traveller's diarrhoea. *The Journal of antimicrobial chemotherapy*, 64(2), 343-347 (2009).
18. Schneider N, Hindle S, Lange G *et al.* Substantial improvements in large-scale redocking and screening using the novel HYDE scoring function. *Journal of Computer-Aided Molecular Design*, 26(6), 701-723 (2011).

19. Bolton EE, Wang Y, Thiessen PA, Bryant SH, Ralph AW, David CS. Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities. In: *Annual Reports in Computational Chemistry*. (Elsevier, 2008) 217-241.
20. Eisenberg D, Luthy R, Bowie JU. VERIFY3D: Assessment of protein models with three-dimensional profiles. In: *Methods in Enzymology*. (Academic Press, 1997) 396-404.
21. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2), 283-291 (1993).
22. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23(4), 566-579 (1995).
23. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 11(9), 739-747 (1998).
24. Bohm HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *Journal of Computer-Aided Molecular Design*, 8(3), 243-256 (1994).
25. Case D, Darden TA, Cheatham TE *et al.* Amber 11. *University of California, San Francisco*, (2010).
26. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3), 712-725 (2006).
27. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2), 926-935 (1983).
28. Davidchack R, Handel R, Tretyakov MV. Langevin thermostat for rigid body dynamics. *The Journal of Chemical Physics*, 130(23), 234101 (2009).
29. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8), 3684-3690 (1984).
30. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3), 327-341 (1977).
31. Darden T, York D, Pedersen L. Particle mesh Ewald: An $N \log(N)$ method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12), 10089-10092 (1993).
32. Amadei A, Linssen AB, HJ B. Essential dynamics of proteins. *Proteins-Structure Function and Bioinformatics*, 17(4), 412-425 (1993).
33. Amadei A, Linssen ABM, de Groot BL, van Aalten DMF, Berendsen HJC. An Efficient Method for Sampling the Essential Subspace of Proteins. *Journal of Biomolecular Structure and Dynamics*, 13(4), 615-625 (1996).
34. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation1. *Journal of Molecular Biology*, 285(4), 1735-1747 (1999).

35. Word JM, Lovell SC, LaBean TH *et al.* Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms1. *Journal of Molecular Biology*, 285(4), 1711-1733 (1999).
36. Doncheva NT, Klein K, Domingues FS, Albrecht M. Analyzing and visualizing residue networks of protein structures. *Trends in Biochemical Sciences*, 36(4), 179-182 (2012).
37. Simossis VA, Heringa J. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Research*, 33(suppl 2), W289-W294 (2005).
38. Ericsson CD, Johnson PC, Dupont HL, Morgan DR, Bitsura JA, de la Cabada FJ. Ciprofloxacin or trimethoprim-sulfamethoxazole as initial therapy for travelers' diarrhea. A placebo-controlled, randomized trial. *Annals of internal medicine*, 106(2), 216-220 (1987).
39. Pazhani GP, Chakraborty S, Fujihara K *et al.* QRDR mutations, efflux system & antimicrobial resistance genes in enterotoxigenic *Escherichia coli* isolated from an outbreak of diarrhoea in Ahmedabad, India. *The Indian journal of medical research*, 134, 214-223 (2011).
40. Berger JM, Fass D, Wang JC, Harrison SC. Structural similarities between topoisomerases that cleave one or both DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, 95(14), 7876-7881 (1998).
41. Humphrey W, Dalke A, K S. VMD - Visual Molecular Dynamics. *J. Molec. Graphics*, 14, 33-38 (1996).
42. Corbett KD, Schoeffler AJ, Thomsen ND, Berger JM. The Structural Basis for Substrate Specificity in DNA Topoisomerase IV. *Journal of Molecular Biology*, 351(3), 545-561 (2005).
43. Saraswathi V, Amit G, Priti H. Intra and Inter-Molecular Communications Through Protein Structure Network. *Current Protein & Peptide Science*, 10(2), 146-160 (2009).
44. Swint-Kruse L. Using Networks To Identify Fine Structural Differences between Functionally Distinct Protein States. *Biochemistry*, 43(34), 10886-10895 (2004).
45. Vendruscolo M, Paci E, Dobson CM, Karplus M. Three key residues form a critical contact network in a protein folding transition state. *Nature*, 409(6820), 641-645 (2001).

CHAPTER 4

Surface Proteome Mining for Identification of Potential Vaccine Candidates against *Campylobacter jejuni*: an *in silico* approach

4.1 Introduction

Campylobacter jejuni are helical-shaped, non-spore forming, microaerophilic gram-negative bacteria and major cause of bacterial campylobacteriosis worldwide [1]. The *Campylobacter* spp. was considered as zoonotic pathogen until isolation of *C. jejuni* was accomplished from human faeces in 1968 [2]. Since its discovery in 1970's, *C. jejuni* remains the most frequent cause of infectious diarrhea attributing to a large economic burden [1]. The first *C. jejuni* (strain NCTC11168) genome was sequenced in 2000 with 94.3% of the genome coding for proteins [3]. *C. jejuni* pathogenesis mechanisms are poorly understood as virulence determinants appear to be multifactorial in nature such as chemotaxis, motility, toxins, flagella, invasion and adherence, surface polysaccharide structures [4]. Antibiotic therapy traditionally involves treatment with erythromycin and ciprofloxacin, but many reports have witnessed resistance of *C. jejuni* to different antibiotics such as tetracycline, kanamycin, chloramphenicol, erythromycin, ciprofloxacin [5,6]. Due to irrational use of antibiotics, antibiotic resistance has escalated posing a challenge to current treatment regimens. Thus, there is a pressing need to develop alternative treatments.

Vaccination has proved cost effective, safest and efficient solution to combat infectious diseases like meningococcal, diphtheria, tetanus, poliomyelitis, pertussis, measles, mumps, rubella in human health care [7]. The traditional approach to subunit vaccine development has negative aspects involving time and labor intensive nature, failure in cases where microorganism cannot be cultured or obtained in sufficient amounts [8]. In order to limit the increasing antibiotic resistance and increasing number of human infections, developing vaccines against *C. jejuni* is both indispensable and attractive. Some mutants of *C. jejuni* with defects in pili or invasion biosynthesis are being evaluated for their protective efficacy in animal models. Flagellin and adhesin proteins have been suggested as potential subunit-based vaccine candidates such as a recombinant truncated flagellin protein (rFla-MBP) conferred 60% protection in a ferret model of diarrhea. Several Killed Whole Cell (WC) and Heat Labile toxin (LT) adjuvanted vaccines are under development [9,10]. In one such example, killed *Campylobacter* whole cell (CWC) organism adjuvanted with heat-labile enterotoxin (LT) of *Escherichia coli* showed protection against intestinal colonisation in mice and rabbits. However, currently there are no approved vaccines available to treat *Campylobacter*-associated illness. Sequencing the genome of many

Campylobacter strains together with development of omics techniques and advanced bioinformatics approaches significantly improve the process of candidate epitopes identification minimizing the arduous peptide screening task for immunobiological properties. The present study has employed a range of computational approaches to investigate the entire proteome of *C. jejuni* for identification of B and T-cell epitopes as potential vaccine candidates. This study has important repercussion for selection of vaccine candidates, a critical step in vaccine development.

4.2 Methods

4.2.1 Retrieving non-homologous proteins from pathogen whole proteome

As described in the workflow diagram (Figure 4.1), the complete proteome of the *C. jejuni* O:2 (strain NCTC 11168) encoding 1,623 proteins was retrieved from Uniprot (Proteome ID UP0000000799). Proteins non-homologous to host from pathogen proteome were retrieved using a two-step filtration procedure. In the first step, sequences with length less than 100 amino acids (aa) were filtered based on the fact that average protein length in bacteria is 267 aa [11]. Consequently sequences with length less than 100 aa would probably not code for any protein. In the next filtration step, sequences were further screened out based on homology with the host (*Homo sapiens*) proteome at an e-value cut off of 0.05. In the BLASTP search, proteins which showed no hits below e-value inclusion threshold were selected as non-homologous pathogen proteins.

4.2.2 Antigenicity and transmembrane prediction

To predict antigenic sequences, these non-homologous pathogen proteins were subjected to VaxiJen server [12], which is based on auto cross covariance (ACC) transformation of protein sequences into uniform vectors of principal amino acid properties with a threshold value of 0.7. The sequences with antigenicity value above threshold were subjected to PSORTb version 3.0 to retrieve outer membrane localized proteins. PSORTb utilizes a Bayesian network model to calculate associated probability for five localization sites viz. cytoplasmic, inner membrane, periplasmic, outer membrane and extracellular with a default probability value (p-value) of 7.5 [13].

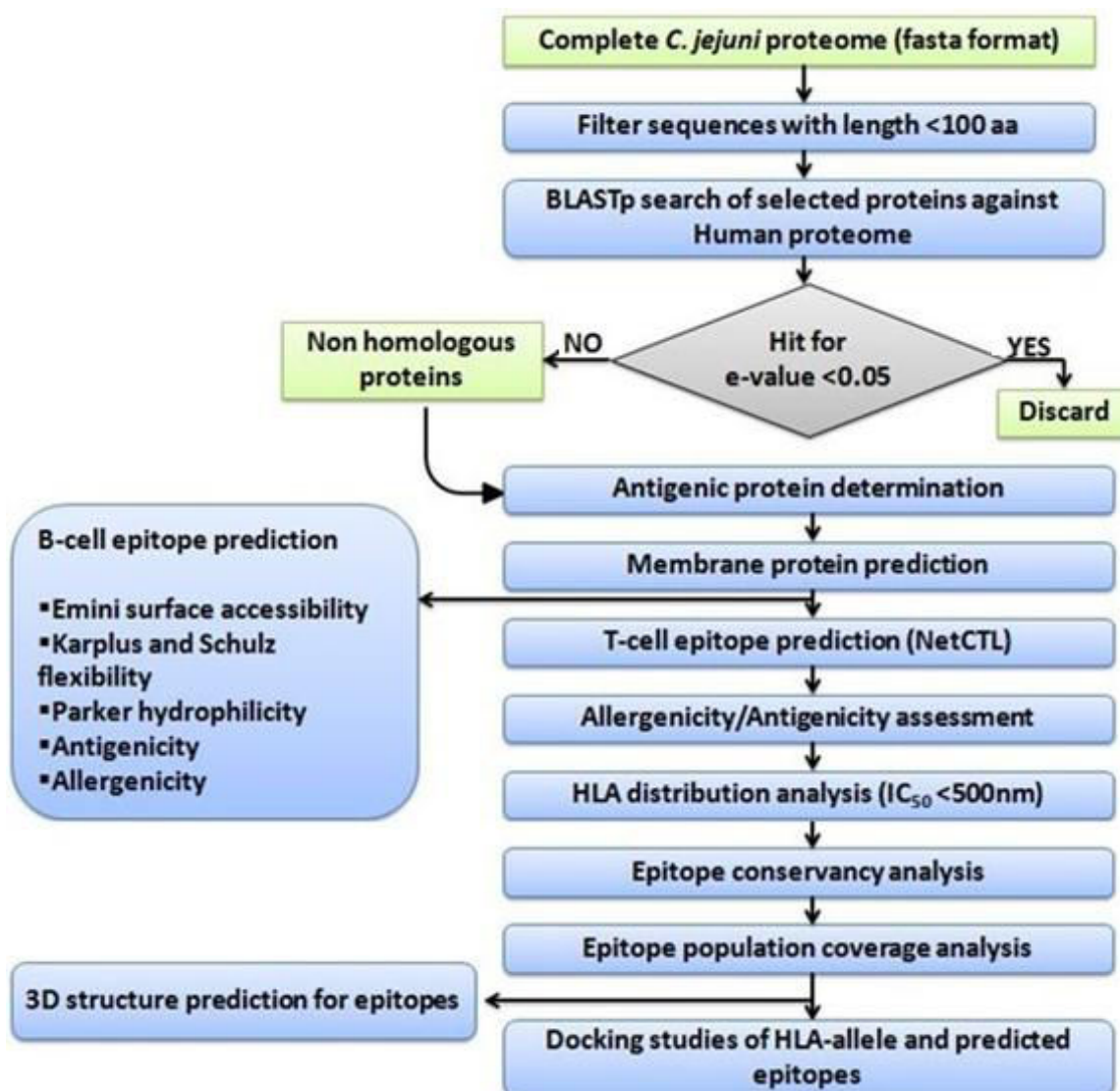


Figure 4.1: Schematic representation of the protocols used for epitope identification.

4.2.3 T-cell epitope prediction

NetCTL 1.2 Server was used to predict Cytotoxic T Lymphocytes (CTL) epitopes from the antigenic sequences localized in outer membrane, at a threshold value of 0.75 to maintain high sensitivity and specificity levels and the prediction was restricted to 12 Major Histocompatibility Class I (MHC-I) supertypes. NetCTL is an artificial neural networks (ANN) and weight matrix based tool combining the prediction of peptide MHC-I binding, proteasomal C terminal cleavage

and TAP transport efficiency [14]. The CTL epitopes generated from NetCTL were assessed for their allergenicity by subjecting them to AllerHunter program which is based on Support Vector Machine (SVM) and pairwise sequence similarity [15]. A threshold value of 0.06 was specified for prediction of cross-reactive allergen.

An Immune Epitope Database (IEDB) tool based on combined predictors of proteasomal processing, TAP transport, and MHC binding was used for predictions of antigen processing through MHC-I [16]. IEDB is the most inclusive database of experimentally characterized B cell and T cell epitopes. The Stabilized Matrix Base Method (SMM) which can model the sequence specificity of quantifiable biological processes [17] was employed to compute Inhibitory Concentration (IC₅₀) values of peptide binding to MHC-I molecules. In conjunction with IEDB tool, MHCpred which uses a partial least squares-based multivariate statistical approach [18] was used for prediction of both MHC-I and MHC-II binders of the predicted peptides. The alleles with binding affinity IC₅₀ value less than 500 nM from both the servers were considered as efficient peptide binders.

4.2.4 Epitope conservancy and HLA-distribution analysis

For each identified peptide, the conservancy was predicted using the IEDB tool [19]. The degree of conservation of each peptide was calculated as the fraction of protein sequences of different strains retrieved from Uniprot that match the aligned peptide sequence above a defined identity level. An IEDB based tool for human population coverage analysis [20] was used to study the distribution of human HLA alleles among the predicted epitopes. The predicted peptides with their corresponding MHC-I and MHC-II alleles were submitted with default parameter settings (the final set containing frequencies of 3,245 alleles for 16 geographical areas, 21 ethnicities, and 115 countries). The predictions were made using the latest dataset from the Allele Frequency Net Database (AFND) [21].

4.2.5 Molecular docking studies of HLA-epitope

4.2.5.1 Designing epitope 3D structure

To study the molecular interactions between the predicted T-cell epitopes (YIQDNFNFY and NTDQAQGTV) and HLA molecules, PEP-FOLD based on a Hidden Markov Model derived

Structural Alphabet (SA) [22] was used to predict the 3D structure of the peptide. PEP-FOLD generated 5 models for input peptide sequence. The best model was selected for docking studies.

4.2.5.2 Docking

To validate our results, we performed a docking study of HLA-A*11:01 and selected epitope using Hex, the first Fourier transform (FFT)-based protein docking server [23]. The crystal structure of HLA-A*11:01 in complex with sars nucleocapsid peptide (PDB Id: 1X7Q) was simplified to HLA-A*11:01, prepared by adding hydrogen atoms. Finally the docking was carried out in Hex using prepared HLA-A*11:01 and PEP-FOLD predicted epitopes as starting structures. The parameters were set to default except for correlation type which uses both shape and electrostatics criteria for docking calculations. The best conformation was selected based on the E_{total} (binding affinity) value and complexes and interactions were visualized in Pymol molecular graphics package [24] and Ligplot respectively [25].

4.2.6 B-cell epitope identification

BCPred [26] and AAP [27] methods at BCPred server, both of which use SVM based classifiers, were utilized with an aim to identify potential antigens which can interact with B lymphocytes. Tools from IEDB were employed to find the B-cell epitopes and further screen out the potential epitopes. Emini-surface accessibility prediction [28], Karplus & Schulz flexibility prediction [29], and Parker hydrophilicity prediction [30] programs were used from IEDB. The regions common to predictions from both BCPred server and IEDB tools were considered as potential B-cell epitopes. These epitopes were further filtered based on allergenicity and antigenicity criteria using AllerHunter and VaxiJen respectively.

4.3. Results

4.3.1 Retrieving non-homologous proteins from pathogen whole proteome

C. jejuni O:2 (strain NCTC 11168) whole proteome encodes 1623 proteins. After filtering out protein sequences on length criteria, we were left with 1500 proteins. We subjected rest of the protein sequences to a homology search against Human proteome database using BLASTP search from standalone blast suite and retrieved a total of 210 pathogen proteins which were non-homologous to humans. Identifying proteins non-homologous to humans is essential as it

excludes the possibility of the peptide vaccine targeting hosts enzymes, thus avoiding adverse effects on humans [31]. Besides, self-peptides can mount an autoimmune response in the host.

4.3.2 Antigenicity and transmembrane prediction

The VaxiJen server used to assess the antigenicity of the protein sequences predicted 157 proteins as antigenic above a threshold of 0.7 which were further analyzed for their cellular location and it revealed that 24 proteins were localized in outer membrane. Identification of outer membrane proteins is critical for reliable and rapid identification of vaccine candidates as many of the vaccines that trigger immune responses appeared to be secreted toxins or surface exposed molecules [32]. Outer membrane localized proteins were further analyzed for vaccine candidate identification.

4.3.3 T-cell epitope prediction

NetCTL predicted T-cell epitopes from each sequence against MHC-supertypes. 28 epitopes with their combinatorial score above threshold 2 were selected from the outer membrane localized antigenic proteins. These epitopes were further assessed by AllerHunter for allergic cross reactivity and by VaxiJen for antigenicity. This step identified four epitopes as potential T-cell epitopes (Table 4.1). For each epitope, SMM based IEDB MHC I processing prediction tool retrieved the MHC-1 alleles with IC50 value less than 500 nM which were potential epitope binders. MHCpred predictions of MHC-I and MHC-II alleles as efficient epitope binders were taken together with IEDB tool predictions to generate a final list of potential binders for each epitope. The results are summarized in Table 4.2.

4.3.4 Epitope conservancy and HLA-distribution analysis

For each predicted epitope, conservancy was determined using IEDB conservancy tool and the results are shown in Table 4.2. Epitope NTDQAQGTV was 75% conserved at identity of >60% while YIQDNFNFY was 50% conserved at 100% identity. Conservancy results for other epitopes (RSDEAQTNY, KSDEEMEKY) were not alluring. Due to scarcity of sequence data in Uniprot database, conservancy results do not portray factual depiction of epitope conservancy. Population coverage analysis was then performed for epitopes NTDQAQGTV and YIQDNFNFY along with their associated MHC-I and MHC-II alleles as input to IEDB

population coverage analysis tool. As shown in Table 4.3, immune response elicitation of the 81.07% and 85.27% world population was covered by the epitopes NTDQAQGTV and YIQDNFNFY respectively. Maximum coverage 85.99% for epitope NTDQAQGTV was in Europe area followed by 85.53%, 84.25% and 80.44% in the population of South Africa, South Asia and North Africa respectively. For epitope YIQDNFNFY, maximum coverage 90.81% was in Europe area followed by South Asia, North America, and Northeast Asia with coverage 85.70%, 84.08%, and 82.80% respectively.

4.3.5 Molecular docking studies of HLA-epitope

Using Hex, different conformations of the predicted epitopes were generated and the best conformation was selected based on binding affinity scores. The docked complexes were visualized in Pymol as shown in Figure 4.2. Figure 4.3 represents the interactions involved in HLA-A*11:01 binding with predicted epitopes. HLA-A*11:01 binds with epitopes NTDQAQGTV, YIQDNFNFY with binding energies -386.53 and -350.09 respectively.

Table 4.1: Most probable predicted epitopes selected on the basis of their NetCTL (MHC binding, proteasomal processing and TAP transport), AllerHunter (Allergic cross-reactivity) and VaxiJen (Antigenicity) Score.

Uniprot Id	Protein Name	Gene Name	Epitope	NetCTL Score	AllerHunter Score	VaxiJen Score
Q0PBW1	Putative TonB-dependent outer membrane receptor	Cj0178	NTDQAQGTV	2.0440	0.04	1.1875
Q0P9M1	Outer membrane component of efflux system (Multidrug efflux system CmeDEF)	cmeD	YIQDNFNFY	2.2174	0.05	0.6824
Q0P9S7	Putative outer-membrane protein	Cj0975	RSDEAQTNV	3.3776	0.06	1.5564
Q0P826	Putative Conserved Protein (Uncharacterized Protein)	Cj1602	KSDEEMEKY	3.3309	0.06	1.1041

Table 4.2: Predicted potential T cell epitopes, along with their interacting MHC-I and MHC-II alleles with an affinity of <500 nM and corresponding IC50 values (bracketed)

Epitope	Total No. of MHC-peptide binders	MHC-I alleles	MHC-II alleles	Conservancy (%)
NTDQAQGTV	24	HLA-A*02:02 (495.45), HLA-A*02:03 (267.30), HLA-A*02:06 (82.22), HLA-A*02:11 (276.27), HLA-A*02:50 (32.01), HLA-A*03:01 (191.87), HLA-A*11:01 (20.61), HLA-A*31:01 (349.14), HLA-A*32:07 (15.78), HLA-A*32:15 (297.34), HLA-A*68:01 (239.88), HLA-A*68:02 (86.10), HLA-A*68:23 (6.57), HLA-A*69:01 (124.99), HLA-B*40:13 (253.47), HLA-C*05:01 (161.41), HLA-C*06:02 (197.61), HLA-C*07:01 (61.12), HLA-C*08:02 (87.33), HLA-C*12:03 (4.22), HLA-C*14:02 (145.08), HLA-C*15:02 (69.09)	HLA-DRB1*01:01 (138.04), HLA-DRB1*04:01 (328.85)	75
YIQDNFNFY	29	HLA-A*01:01 (3.94), HLA-A*02:02 (85.70), HLA-A*02:03 (126.47), HLA-A*02:06 (16.98), HLA-A*02:17 (190.84), HLA-A*03:01 (69.18), HLA-A*11:01 (69.82), HLA-A*25:01 (296.54), HLA-A*26:01 (306.63), HLA-A*26:02 (478.98), HLA-A*29:02 (49.29), HLA-A*30:02 (228.80), HLA-A*32:07 (41.50), HLA-A*32:15 (95.55), HLA-A*68:01 (232.27), HLA-A*68:23 (27.46), HLA-A*80:01 (87.30), HLA-B*15:01 (123.59), HLA-B*15:02 (94.02), HLA-B*15:03 (413.13), HLA-B*27:20 (134.12), HLA-B*35:01 (184.75), HLA-B*40:13 (138.33), HLA-C*03:03 (25.62), HLA-C*12:03 (13.59), HLA-C*14:02 (124.34)	HLA-DRB1*01:01 (0.60), HLA-DRB1*04:01 (158.49), HLA-DRB1*07:01 (144.21)	50
RSDEAQTNY	21	HLA-A*01:01 (108.39), HLA-A*11:01 (148.94), HLA-A*30:02 (245.73), HLA-A*31:01 (431.52), HLA-A*32:07 (11.91), HLA-A*32:15 (196.45), HLA-A*68:01 (367.28), HLA-A*68:23 (17.81), HLA-B*15:02 (117.06), HLA-B*15:02 (401.08), HLA-B*15:03 (153.85), HLA-B*15:17 (60.63), HLA-B*27:20 (6.48), HLA-B*40:13 (40.64), HLA-B*58:01 (415.24), HLA-C*05:01 (161.41), HLA-C*06:02 (275.30), HLA-C*07:01 (121.39), HLA-C*08:02 (388.29), HLA-C*12:03 (14.81)	HLA-DRB1*01:01 (8.07)	10.34
KSDEEMEKY	19	HLA-A*01:01 (98.40), HLA-A*02:02 (119.40), HLA-A*02:03 (98.17), HLA-A*02:06 (232.27), HLA-A*11:01 (144.54), HLA-A*30:02 (328.44), HLA-A*32:07 (18.80), HLA-A*32:15 (162.65), HLA-A*68:23 (25.99), HLA-B*15:17 (222.69), HLA-B*27:20 (5.57), HLA-B*40:13 (25.88), HLA-B*58:01 (363.33), HLA-C*05:01 (86.83), HLA-C*07:01 (491.10), HLA-C*12:03 (4.95), HLA-C*14:02 (315.22)	HLA-DRB1*01:01 (34.20), HLA-DRB1*07:01 (259.42)	100

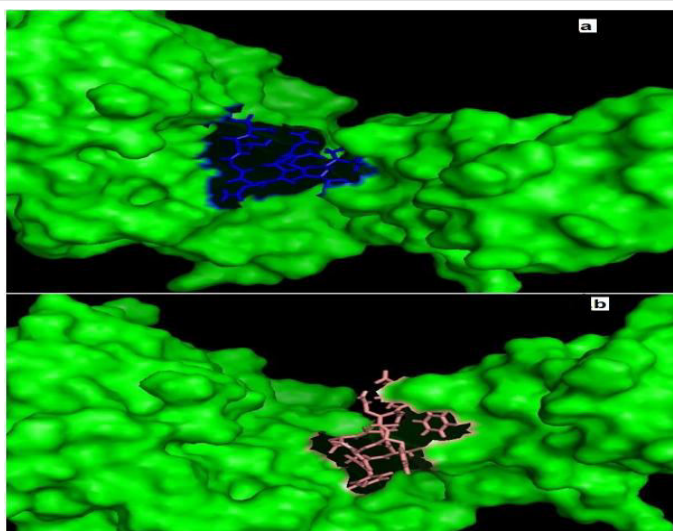


Figure 4.2: Docked complexes of HLA-A*11:01 against predicted epitopes generated by Hex docking program. a) Epitope NTDQAQGTV b) Epitope YIQDNFNFY.

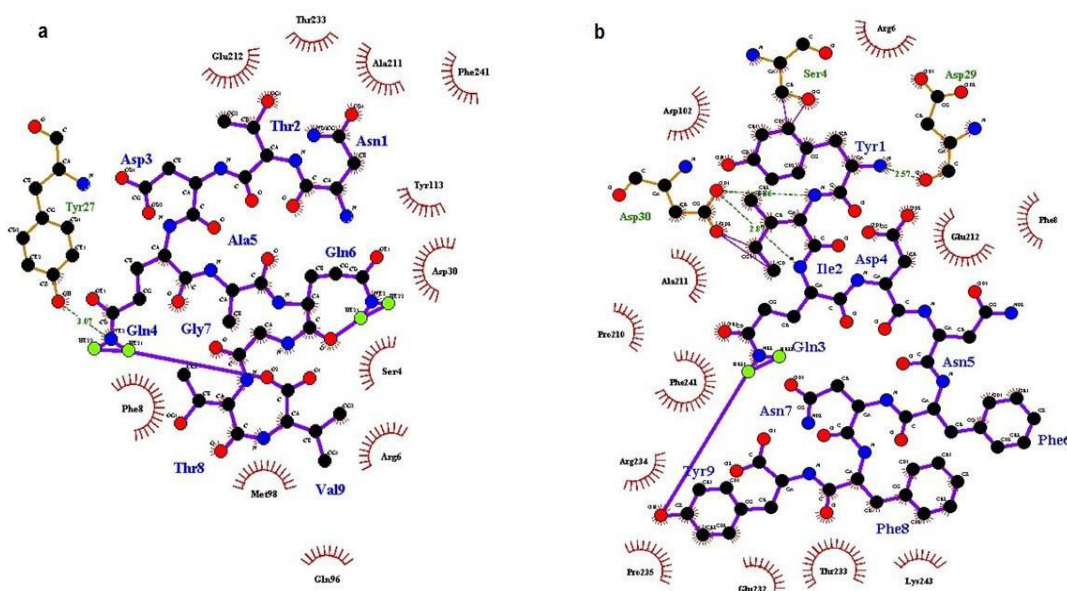


Figure 4.3: Interactions involved in HLA-A*11:01 binding with a) Epitope NTDQAQGTV b) Epitope YIQDNFNFY.

4.3.6 B-cell epitope identification

As per the criteria set for prediction of B-cell epitopes, Table 4.4 depicts the epitopes predicted using AAP, BCPred and IEDB tools further filtered based on allergenicity and antigenicity properties. Antigenic regions common to both BCPred and AAP were subjected to IEDB Emini

surface accessibility tool to predict peptides which were surface exposed. Predicted peptides were checked for flexibility and hydrophilicity using IEDB tools Karplus & Schulz flexibility prediction, and Parker hydrophilicity prediction. This yielded a total of 25 peptides as B-cell epitopes. To test such peptides for their potential as B-cell epitopes, they were checked for their allergenicity and antigenicity which yielded four epitopes with allergenicity score ≤ 0.06 and antigenicity score > 1 as shown in Table 4.4.

4.4 Discussion

With the advancement in sequencing technologies, there has been remarkable progress in the vaccinology area, enabling researchers to finally move beyond the traditional vaccinology approach. With computational approaches it is now feasible to access the entire antigenic repertoire of an organism. Reverse Vaccinology (RV) approach to vaccine identification came into existence with addressing the problem of vaccine identification against *Meningococcus B* (Men B). Men B is a pathogen which was intractable to vaccine development using conventional vaccinology approach as its capsular polysaccharide is identical to a human self-antigen [33]. Hitherto, RV has been practically applied against many pathogens [34,35]. In the post-genomic era, power of omics data has been complemented by bioinformatics approaches which may lead to the discovery of unique antigens that may eventually improve existing vaccines. Many researchers have already proposed an epitope-based vaccine candidate against *C. jejuni* with their studies aimed at identifying vaccine candidates from specific proteins like cytolethal distending toxin (CDT), autotransporter protein CapA, polysaccharide capsules etc. [36-38]. Developing killed WC vaccines is complicated by dearth of information on pathogenesis of *C. jejuni* and development of flagellin subunit-based vaccines is complicated owing to antigenic diversity of *Campylobacter* flagellins. Perceiving the gaps in current efforts for vaccine development against *Campylobacter*, we have undertaken current study of genome wide screening of *C. jejuni* using an *in silico* approach, aimed at identifying potential vaccine candidates against this organism and expedite the efforts in this direction. Currently most vaccines are based on B-cell providing antibody mediated immunity. However, T –cells confer long-lasting immunity while antibody mediated immunity can be easily overcome by surge of antigens [39]. Cytotoxic CD8⁺ T lymphocytes (CTL) hamper infectious agents from spreading by invading infected cells.

Table 4.3: Population coverage of predicted epitopes based on MHC-I and MHC-II restriction data. **4.3a)** for epitope NTDQAQGTV maximum population coverage by Europe **4.3b)** for epitope YIQDNFNFY maximum population coverage by Europe

Population / Area	Class I and II Coverage	Population / Area	Class I and II Coverage
World	81.07%	World	85.27%
East Asia	68.22%	East Asia	80.61%
Northeast Asia	74.14%	Northeast Asia	82.80%
South Asia	84.25%	South Asia	85.70%
Southeast Asia	63.61%	Southeast Asia	70.72%
Southwest Asia	72.88%	Southwest Asia	70.55%
Europe	85.99%	Europe	90.81%
East Africa	70.98%	East Africa	66.55%
West Africa	69.64%	West Africa	75.91%
Central Africa	75.76%	Central Africa	69.61%
North Africa	80.44%	North Africa	77.05%
South Africa	85.53%	South Africa	77.44%
West Indies	46.39%	West Indies	77.88%
North America	78.72%	North America	84.08%
Central America	1.34%	Central America	15.10%
South America	69.25%	South America	59.04%
Oceania	63.73%	Oceania	64.30%

4.3a**4.3b**

Thus, in this study we have proposed both B and T-cell epitopes which could be experimentally tested for their efficacy in triggering humoral and cell mediated immune responses. As described in schematic workflow diagram (Figure 4.1), we have framed a set of criteria for identifying potential vaccine candidates which involves antigenicity, T-cell/B-cell processivity, interaction with HLA alleles, allergenicity, conservancy, and population coverage. Protective epitopes are not clearly defined for *C. jejuni*. Thus, while screening proteomic data, it is of utmost importance to select the proteins which can confer protection. To select such segments from the proteins, it is encouraged to select genomic segments with antigenic properties. Thus, antigenicity filter was employed at several stages of vaccine candidate identification task. Initially, the proteins with antigenicity score above threshold 0.7 were selected as antigenic. Identified B and T-cell epitopes were also filtered based on antigenicity criterion.

Physiochemical properties like flexibility, hydrophilicity, solvent accessibility are distinctive features of B-cell epitopes. These features have been exploited in many B-cell epitope prediction programs [40]. Initially, based on surface accessibility, flexibility and hydrophilicity criteria B-cell epitopes which could be proficiently processed by B-lymphocytes were identified. NetCTL server predicted T-cell epitopes based on combined predictions of MHC class I binding, proteasomal C terminal cleavage and TAP transport efficiency. *C. jejuni* strains are highly diverse which further complicates the vaccine development against this pathogen. Consequently conservation of the epitopes at sequence level reveals that these regions are imperative from evolutionary point. Population coverage plays an essential role in vaccine development process. Our predicted peptides showed good population coverage in spite of the fact that in case of MHC-II data was only available for the alleles HLA-DRB1*01:01, HLA-DRB1*04:01, and HLA-DRB1*07:01. Though, all the predicted nonamers were interacting with most common HLA allele HLA-DRB1*01:01 as shown in Table 4.2. For the predicted epitopes, in developing world highest population coverage was in Asian and African countries where the diarrheal incidence rate is reported to be highest and in industrialized nations, it was highest for Europe and North America, aligning with the fact that maximum number of travelers to Asia and Africa are observed from these countries [41].

Further investigation in the data shows that epitope NTDQAQGTV has high antigenicity value (Table 4.2) but epitope YIQDNFNFY has a maximum of 29 MHC-interacting alleles. Epitope YIQDNFNFY was identified from *cmeD* which encodes for outer membrane component of Multidrug efflux system CmeDEF. In a study, CmeC which is an essential outer membrane component of CmeABC multidrug efflux pump was proposed as a promising subunit vaccine candidate against *C. jejuni* infection using chicken model [42]. *cmeDEF* also plays important role in antibiotic resistance against several antibiotics and toxic compounds. *cmeABC* and *cmeDEF* act synergistically in retaining cell viability and conferring antibiotic resistance [43]. Epitope sequence NTDQAQGTV has lower IC_{50} value of 20.61 nM for MHC supertype (HLA-A*11:01) as compared to YIQDNFNFY which has an IC_{50} of 69.82 nM with HLA-A*11:01. The results of computational docking studies coincide with the binding affinity values. NTDQAQGTV has a stronger affinity for HLA-A*11:01 with binding energy of -386.53 while YIQDNFNFY binds in the groove of HLA-A*11:01 with total energy -350.09. Epitope sequence YIQDNFNFY is more conserved at 100% identity and has a high score as a processed peptide as

evidenced from NetCTL score (Table 4.1). As seen in Table 4.3, population coverage analysis reveals that epitope YIQDNFNFY covers a large proportion of human population. Lowest coverage for this epitopes is in Central America (15.10%) which is much higher when compared to population coverage of NTDQAQGTV in the same region being 1.34%.

Table 4.4: Four most potential B-cell epitopes by combined predictions of AAP, BCPred, IEDB tools (Emini Surface Accessibility, Karplus and Schulz flexibility, Parker hydrophilicity), filtered based on their AllerHunter and VaxiJen score.

Uniprot Id	Protein Name	Gene Name	Start	End	B-cell Epitope	Length	Aller-Hunter Score	VaxiJen Score
Q0PBW1	Putative TonB-dependent outer membrane receptor	Cj0178	648	660	YTGKAKRVNPNT	12	0.06	1.6624
Q0PAY6	Putative periplasmic protein	Cj0530	583	593	IYRKHSNSSNS	11	0.06	1.6432
Q0P9M1	Outer membrane component of efflux system (Multidrug efflux system CmeDEF)	cmeD	220	228	NPQQEKSQN	9	0.05	1.2950
			327	336	RFSEKKNKEE	10	0.06	1.3154

Four potential B-cell epitopes were predicted with their VaxiJen score (antigenicity) >1 and with AllerHunter score (allergenicity) ≤ 0.06 threshold. The epitopic sequence YTGKAKRVNPNT has highest antigenicity of 1.6624 followed by IYRKHSNSSNS and RFSEKKNKEE with antigenicity scores 1.6432 and 1.3154 respectively. Based on AllerHunter results, epitope sequence NPQQEKSQN has highest possibility of being a non-allergen as marked by the lowest score 0.05, while other B-cell epitopes YTGKAKRVNPNT, IYRKHSNSSNS and RFSEKKNKEE have AllerHunter score of 0.06. Elicitation of effective immune responses depends on the specificity and diversity of the T-cell epitopes binding to HLA-alleles. Due to highly polymorphic nature of MHC, it is desirable to identify peptides which can bind to many MHC-alleles [44]. Our predicted T-cell epitopes NTDQAQGTV and YIQDNFNFY bind to more

than 20 MHC-alleles and have broad human population coverage. HLA-A*11:01 was selected for docking studies. The predicted T cell epitopes interact with HLA-A*11:01 with varying affinities. From computational docking results, it was interpreted that the epitopes bind efficiently to the HLA-A*11:01. It is believed that such a systematic computational pipeline for prediction of vaccine candidates when employed to *C. jejuni* proteome reveals epitopes that would be able to elicit an efficacious immune response. There has been growing body of evidence which state the indispensable role of bioinformatics approaches in translational medicine [36-38]. As shown in Figure 4.4, there was sturdy decline in the protein search space at each step. It was noticed that there was significant reduction in the proteome size to be searched for vaccine identification. We started with the proteome size of 1623 proteins. Applying different filtration criteria at each step, we were left with 8 antigenic sequences which were eventually tested for presence of vaccine candidates. In summary, omics-guided approaches and bioinformatics analyses offer broad potential for further developments in global health relevant novel therapeutics.

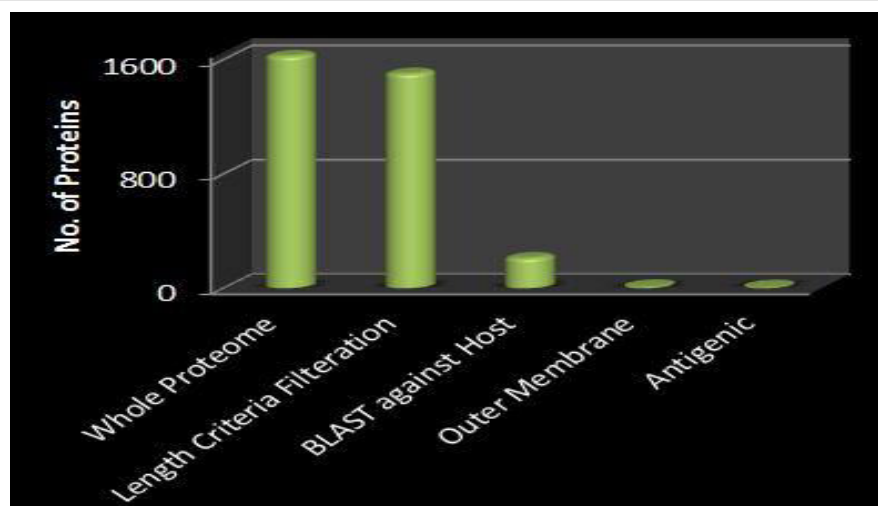


Figure 4.4: Step-wise reduction in the total no of proteins in search for the identification of vaccine candidates against *C. jejuni*.

Conclusion

Traditional molecular immunology techniques for vaccine identification are time and labor consuming. A wide array of omics techniques, whole genome sequencing data and novel

bioinformatics approaches have substantially improved our systemic understanding of complex diseases. These techniques hold a greater potential to be utilized for rapid and reliable genome wide screening for identification of vaccine candidates; thus have hastened the pace of vaccine development to a great extent by significantly reducing the number of experimentally testable epitopes. Our predicted epitopes are prospective vaccine candidates on grounds of higher population coverage and interactions with many HLA alleles. In conclusion, immunoinformatics-based approach was utilized for detection of protective antigens in *C. jejuni*, which may serve as potential vaccine candidates to control campylobacteriosis once validated experimentally *in vitro* and *in vivo*. This immunoinformatics-based approach can be applied to other hosts or other enteric pathogens.

References

1. Friedman CR, Neimann J, Wegener HC, Tauxe RV. Epidemiology of *Campylobacter jejuni* infections in the United States and other industrialized nations. In: *Campylobacter*. (ASM International, Washington, USA, 2000) 121-138.
2. Dekeyser P, Gossuin-Detrain M, Butzler JP, Sternon J. Acute Enteritis Due to Related *Vibrio*: First Positive Stool Cultures. *Journal of Infectious Diseases*, 125(4), 390-392 (1972).
3. Parkhill J, Wren BW, Mungall K *et al*. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, 403(6770), 665-668 (2000).
4. Ketley JM. Pathogenesis of Enteric Infection by *Campylobacter*. *Microbiology*, 143(1), 5-21 (1997).
5. Alfredson DA, Korolik V. Antibiotic resistance and resistance mechanisms in *Campylobacter jejuni* and *Campylobacter coli*. *FEMS Microbiology Letters*, 277(2), 123-132 (2007).
6. Thakur S, Zhao S, McDermott PF *et al*. Antimicrobial Resistance, Virulence, and Genotypic Profile Comparison of *Campylobacter jejuni* and *Campylobacter coli* Isolated from Humans and Retail Meats. *Foodborne Pathogens and Disease*, 7(7), 835-844 (2010).
7. Moriel DG, Scarselli M, Serino L, Mora M, Rappuoli R, Maignani V. Genome-based vaccine development: A short cut for the future. *Human Vaccines*, 4(3), 184-188 (2008).
8. Rinaudo CD, Telford JL, Rappuoli R, Seib KL. Vaccinology in the genome era. *The Journal of Clinical Investigation*, 119(9), 2515-2525 (2009).
9. O'Ryan M, Vidal R, del Canto F, Carlos Salazar J, Montero D. Vaccines for viral and bacterial pathogens causing acute gastroenteritis: Part II: Vaccines for *Shigella*, *Salmonella*, enterotoxigenic *E. coli* (ETEC) enterohemorrhagic *E. coli* (EHEC) and *Campylobacter jejuni*. *Human Vaccines & Immunotherapeutics*, 11(3), 601-619 (2015).
10. Albert MJ. Vaccines against *Campylobacter jejuni*. *Austin Journal of Clinical Immunology*, 1(3), 1013 (2014).

11. Brocchieri L, Karlin S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, 33(10), 3390-3400 (2005).
12. Doytchinova I, Flower D. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, 8(1), 4 (2007).
13. Yu NY, Wagner JR, Laird MR *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26(13), 1608-1615 (2010).
14. Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics*, 8, 424-424 (2007).
15. Muh HC, Tong JC, Tammi MT. AllerHunter: A SVM-Pairwise System for Assessment of Allergenicity and Allergic Cross-Reactivity in Proteins. *PLoS ONE*, 4(6), e5861 (2009).
16. Tenzer S, Peters B, Bulik S *et al.* Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cellular and Molecular Life Sciences CMLS*, 62(9), 1025-1037 (2005).
17. Peters B, Sette A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*, 6, 132-132 (2005).
18. Guan P, Doytchinova IA, Zyngouri C, Flower DR. MHCpred: a server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Research*, 31(13), 3621-3624 (2003).
19. Bui H-H, Sidney J, Li W, Fusseder N, Sette A. Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. *BMC Bioinformatics*, 8, 361-361 (2007).
20. Bui H-H, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics*, 7, 153-153 (2006).
21. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Research*, 39(suppl 1), D913-D919 (2011).
22. Thevenet P, Shen Y, Maupetit J, Guyon Fdr, Derreumaux P, Tuffery P. PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic Acids Research*, 40(Web Server issue), W288-W293.
23. Macindoe G, Mavridis L, Venkatraman V, Devignes M-D, Ritchie DW. HexServer: an FFT-based protein docking server powered by graphics processors. *Nucleic Acids Research*, 38(Web Server issue), W445-W449 (2010).
24. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.3r1. (2010)
25. Laskowski RA, Swindells MB. LigPlot+: Multiple Ligand-Protein Interaction Diagrams for Drug Discovery. *Journal of Chemical Information and Modeling*, 51(10), 2778-2786 (2011).
26. El-Manzalawy Y, Dobbs D, Honavar V. Predicting linear B-cell epitopes using string kernels. *Journal of molecular recognition : JMR*, 21(4), 243-255 (2008).

27. Chen J, Liu H, Yang J, Chou KC. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids*, 33(3), 423-428 (2007).
28. Emini EA, Hughes JV, Perlow DS, Boger J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *Journal of Virology*, 55(3), 836-839 (1985).
29. Karplus PA, Schulz GE. Prediction of chain flexibility in proteins. *Naturwissenschaften*, 72(4), 212-213 (1985).
30. Parker JMR, Guo D, Hodges RS. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites. *Biochemistry*, 25(19), 5425-5432 (1986).
31. Butt AM, Nasrullah I, Tahir S, Tong Y. Comparative Genomics Analysis of *Mycobacterium ulcerans* for the Identification of Putative Essential Genes and Therapeutic Candidates. *PLoS ONE*, 7(8), e43080 (2012).
32. Doro F, Liberatori S, Rodriguez-Ortega MJ *et al.* Surfome Analysis as a Fast Track to Vaccine Discovery: identification of a novel protective antigen for group B *Streptococcus* hypervirulent strain COH1. *Molecular & Cellular Proteomics : MCP*, 8(7), 1728-1737 (2009).
33. Giuliani MM, Adu-Bobie J, Comanducci M *et al.* A universal vaccine for serogroup B meningococcus. *Proceedings of the National Academy of Sciences of the United States of America*, 103(29), 10834-10839 (2006).
34. Maione D, Margarit I, Rinaudo CD *et al.* Identification of a Universal Group B *Streptococcus* Vaccine by Multiple Genome Screen. *Science (New York)*, 309(5731), 148-150 (2005).
35. Thorpe C, Edwards L, Snelgrove R *et al.* Discovery of a vaccine antigen that protects mice from *Chlamydia pneumoniae* infection. *Vaccine*, 25(12), 2252-2260 (2007).
36. Ingale A, Goto S. Prediction of CTL epitope, in silico modeling and functional analysis of cytolethal distending toxin (CDT) protein of *Campylobacter jejuni*. *BMC Research Notes*, 7(1), 92 (2014).
37. Ashgar SSA, Oldfield NJ, Wooldridge KG *et al.* CapA, an Autotransporter Protein of *Campylobacter jejuni*, Mediates Association with Human Epithelial Cells and Colonization of the Chicken Gut. *Journal of Bacteriology*, 189(5), 1856-1865 (2007).
38. Guerry P, Poly F, Riddle M, Maue AC, Chen Y-H, Monteiro MA. *Campylobacter* polysaccharide capsules: virulence and vaccines. *Frontiers in Cellular and Infection Microbiology*, 2, 7 (2012).
39. Bacchetta R, Gregori S, Roncarolo M-G. CD4+ regulatory T cells: Mechanisms of induction and effector function. *Autoimmunity Reviews*, 4(8), 491-496 (2005).
40. Li X, Yang H-W, Chen H, Wu J, Liu Y, Wei J-F. In Silico Prediction of T and B Cell Epitopes of Der f 25 in *Dermatophagoides farinae*. *International Journal of Genomics*, 2014, 10 (2014).
41. Harris JA, Roy K, Woo-Raspberry V *et al.* Directed Evaluation of Enterotoxigenic *Escherichia coli* Autotransporter Proteins as Putative Vaccine Candidates. *PLoS Neglected Tropical Diseases*, 5(12), e1428 (2011).
42. Zeng X, Xu F, Lin J. Development and Evaluation of CmeC Subunit Vaccine against *Campylobacter jejuni*. *Journal of vaccines & vaccination*, 1(3), 112 (2010).

43. Akiba M, Lin J, Barton Y-W, Zhang Q. Interaction of CmeABC and CmeDEF in conferring antimicrobial resistance and maintaining cell viability in *Campylobacter jejuni*. *Journal of Antimicrobial Chemotherapy*, 57(1), 52-60 (2006).
44. Germain RN. MHC-dependent antigen processing and peptide presentation: Providing ligands for T lymphocyte activation. *Cell*, 76(2), 287-299 (1994).

CHAPTER 5

Identification of Epitope-Based Peptide Vaccine Candidates against Enterotoxigenic *Escherichia coli*: A Comparative Genomics and Immunoinformatics Approach

5.1 Introduction

Enterotoxigenic *Escherichia coli* (ETEC), a gram negative, rod shaped bacilli of Enterobacteriaceae family colonizing the mucosal surface of small intestine, is the leading cause of bacterial diarrhea in the endemic countries and major cause of Traveler's Diarrhea (TD) worldwide. Diarrhea due to ETEC remains a serious problem with an estimated annual incidence rate of 840 million infections and 3,80,000 deaths worldwide [1]. Beyond the startling mortality rates, are the emerging evidences for long-term consequences of diarrheal disease-associated intestinal enteropathy and malnutrition on physical and cognitive development [2]. ETEC's major virulence mechanism involves production of one or both of enterotoxins, heat-labile enterotoxin (LT) and heat-stable enterotoxin (ST) and many colonization factors (pili/fimbrial or nonfimbrial) [3], 23 of which are defined while many still remain uncharacterized [4]. ETEC is difficult to recognize as it is not readily detected by standard assays and is often underestimated as being a major diarrheal causative agent.

Vaccines have proved a breakthrough in disease prevention ultimately improving life expectancy against diseases like meningococcal, diphtheria, tetanus, poliomyelitis, pertussis, measles, mumps, rubella [5]. Although conventional vaccinology strategies have been relatively successful, they are inadequate in vaccine development against antigenically diverse pathogens and in some cases conventional approaches fail due to their low efficacy and safety issues [6]. The leading cellular vaccine candidates against ETEC include inactivated and live attenuated ETVAX and ACE527 vaccines respectively. ACE527 is a three-strain-combination live attenuated whole cell vaccine while ETVAX consists of four inactivated recombinant *E. coli* strains overexpressing the major ETEC colonization factors mixed with an LT B-subunit related toxoid. ETEC vaccine candidates based on subunits, toxins, or novel antigens such as new subunit ETEC candidate using the tip proteins from fimbriae, a multicomponent fusion protein, dmLT, Cholera whole cell/B subunit, ST constructs, YghJ (type 2 secretion system effector molecule), EatA (serine protease) etc. are also being evaluated for their efficacy [7]. Hitherto there are no licensed vaccines available against ETEC.

With the sequencing of first bacterial genome of *Haemophilus influenza* in 1995 [8], vaccine development entered a new era where it was feasible to access the entire antigenic repertoire of an organism with the information present in its genome sequence, an approach designated as

Reverse Vaccinology (RV). Since the first successful application of RV to *Neisseria meningitidis* serogroup B (MenB) [9], there have been significant improvements to classical RV approach and it has been successfully applied to a broad spectrum of pathogens [10-12]. Application of omics techniques together with advanced bioinformatics tools have significantly improved the process of candidate epitopes identification with potential to be exploited as vaccine candidates; thus minimizing the arduous screening task for peptides with immunobiological properties. The present study was undertaken to design an epitope based vaccine against ETEC using the RV approach which work through the identification of peptides capable of binding to human leukocyte antigens (HLA) and being processed by T cells. A range of computational approaches were employed to probe the entire antigenic repertoire of ETEC for identification of potential B and T-cell epitopes. This study has important reverberations for selection of vaccine candidates against ETEC which will eventually reduce the impact of disease caused by ETEC infections.

5.2 Methods

5.2.1 Comparative genomics analysis

As illustrated in Figure 5.1, a two step comparative genomics approach was developed to retrieve orthologous proteins of *E. coli* O139:H28 (strain E24377A / ETEC) with other pathogenic strains and excluded orthologous proteins from other commensal/nonpathogenic *E. coli* strains. To achieve this, we included two pathogenic ETEC strains (E24377A and H10407) and one commensal strain (*E. coli* SE11). The complete proteomes of the diarrheagenic *E. coli* O139:H28 (strain E24377A/ETEC), *E. coli* O78:H11 (strain H10407/ETEC) and commensal *E. coli* (strain SE11) were downloaded from the Uniprot database with Proteome Ids UP000001122, UP000006877 and UP000008199 respectively. The most probable set of orthologous proteins shared by the two pathogenic ETEC strains were identified using a reciprocal best-hit criterion as described. In the first step, entire proteome of one strain was BLASTP searched against the proteome of other strain using a threshold E value of 10^{-12} . To scrutinize orthologous proteins, the alignment region between the subject and the query protein had to be at least 80% of the entire protein length, and similarity at least 40% for both query and target sizes. In the second step, the obtained orthologous proteins were again searched against commensal strain using BLASTP at a cut-off E value of 10^{-12} . Proteins shared between pathogenic and commensal strain

were excluded.

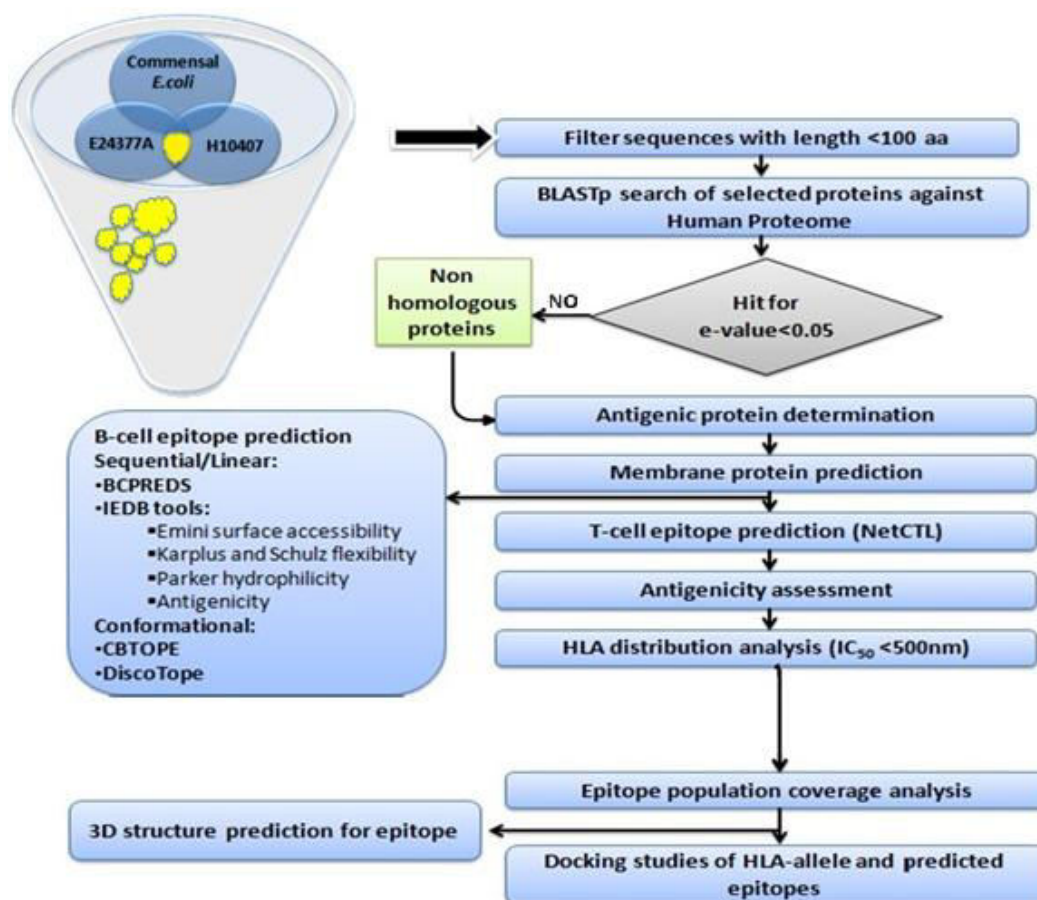


Figure 5.1: Schematic representation of the steps involved in target protein prioritization and antigenic epitope mapping in ETEC E24377A strain using an integrated comparative genomics and immunoinformatics approach.

The orthologous proteins with length less than 100 amino acid (aa) were filtered as the average protein length in bacteria is 267 aa [13]. Therefore proteins with length less than 100 aa are unlikely to code for any protein. In the next step, sequences were further screened out based on homology with the host (*Homo sapiens*) proteome at an e-value cut off of 0.05. In the BLASTP search, proteins which showed no hits below e-value inclusion threshold were selected as non-homologous pathogen proteins.

5.2.2 Antigenicity and transmembrane prediction

Non-homologous pathogen proteins were subjected to VaxiJen server [14], which is based on auto cross covariance (ACC) transformation of protein sequences into uniform vectors of principal amino acid properties, with threshold value 0.7 for prediction of antigenic sequences. The sequences with antigenicity value above threshold were scrutinized for their cellular location by subjecting them to PSORTb version 3.0. A probability value (p-value) of 7.5 was set to predict the antigenic sequences localized in membrane. PSORTb utilizes a Bayesian network model to calculate associated probability for each localization site [15].

5.2.3 T-cell epitope prediction

NetCTL 1.2 Server was used to predict Cytotoxic T Lymphocytes (CTL) epitopes from the antigenic sequences localized in membrane, at a threshold value of 0.75 to maintain high sensitivity and specificity levels and the prediction was restricted to 12 Major Histocompatibility Class I (MHC-I) supertypes. NetCTL is an artificial neural networks (ANN) and weight matrix based tool combining the prediction of peptide MHC-I binding, proteasomal C terminal cleavage and TAP transport efficiency [16]. The CTL epitopes generated from NetCTL were assessed for their antigenicity by subjecting them to VaxiJen server. Identified epitopes were categorized into low, medium and high priority vaccine candidates based upon the VaxiJen scores. An Immune Epitope Database (IEDB) tool based on combined predictors of proteasomal processing, TAP transport, and MHC binding was used for predictions of antigen processing through MHC-I [17]. IEDB is the most inclusive database of experimentally characterized B cell and T cell epitopes. The Stabilized Matrix Base Method (SMM) which can model the sequence specificity of quantifiable biological processes [18] was employed to compute Inhibitory Concentration (IC₅₀) values of peptide binding to MHC-I molecules. In conjunction with IEDB tool, MHCpred which uses a partial least squares-based multivariate statistical approach [19] was used for prediction of both MHC-I and MHC-II binders of the predicted peptides. The alleles with binding affinity IC₅₀ value less than 500 nM from both the servers were considered as efficient peptide binders.

5.2.4 HLA-distribution analysis

An IEDB based tool for human population coverage analysis [20] was used to study the distribution of human HLA alleles among the predicted epitopes. Human population coverage analysis tool was input with the predicted peptides and their corresponding MHC-I and MHC-II

alleles. All other parameters were set to default with the final data set containing frequencies of 3,245 alleles for 16 geographical areas, 21 ethnicities, and 115 countries). The predictions were made using the latest dataset from the Allele Frequency Net Database (AFND) [21].

5.2.5 Molecular docking studies of HLA-epitope

5.2.5.1 Designing epitope 3D structure

To study the molecular interactions between the predicted T-cell epitope (SSFASNNIY) and HLA molecules, PEP-FOLD based on a Hidden Markov Model derived Structural Alphabet (SA) [22] was used to predict the 3D structure of the peptide. PEP-FOLD generates 5 models for input peptide sequence, the best of which is used for docking studies.

5.2.5.2 Docking

To validate our results, we performed a docking study of HLA-A*11:01 and high priority epitopes using Hex, the first Fourier transform (FFT)-based protein docking server [23]. The crystal structure of HLA-A*11:01 in complex with sars nucleocapsid peptide (PDB Id: 1X7Q) was simplified to HLA-A*11:01, prepared by adding hydrogen atoms. Finally the docking was carried out in Hex using prepared HLA-A*11:01 and PEP-FOLD predicted epitopes as starting structures. The parameters were set to default except for correlation type which uses both shape and electrostatics criteria for docking calculations. Hex uses spherical polar Fourier (SPF) algorithm for rotational correlations and abridged execution times. Hex resembles molecules to 3D parametric functions relating electrostatic charge, surface shapes, and potential dissemination which engage electrostatic and van der Waals interactions. The best conformation was selected based on the E_{total} (binding affinity) value. The docked complexes and interactions were visualized in Pymol molecular graphics package [24] and Ligplot respectively [25].

5.2.6 B-cell epitope identification

Both sequential and conformational B-cell epitopes were predicted. To predict sequential B-cell epitopes, BCPred [26] and AAP [27] algorithms at BCPREDS server, both of which use SVM based classifiers, were utilized with an aim to identify potential antigens which can interact with B lymphocytes. Tools from IEDB were employed to find the B-cell epitopes and further screen out the potential epitopes. Emini-surface accessibility prediction [28], Karplus & Schulz

flexibility prediction [29], and Parker hydrophilicity prediction [30] programs were used from IEDB. The regions common to predictions from both BCPREDS server and IEDB tools were considered as potential B-cell epitopes. These epitopes were further scrutinized for antigenicity prediction using VaxiJen. For prediction of conformational B-cell epitopes, CBTOPE server was used which uses an SVM-based model to predict the probability of a residue being a part of a B-cell epitope [31]. DiscoTope server was also used for conformational B-cell epitopes prediction from protein 3D structures; which is based on surface accessibility calculation for assigning a propensity score to each residue [32]. To predict 3D structures of the outer membrane proteins to be used in DiscoTope Server, I-TASSER was used which is an integrated platform for automated protein structure and function prediction [33].

5.3 Results

5.3.1 Genomic Features

The chosen strain, ETEC E24377A (Uniprot Id: UP000001122) is composed of 4.9 Mbp long circular genome having 4,773 coding genes. This rod shaped facultatively anaerobic chemoorganotrophic gram-negative bacterium has about 50.6 % GC content (Figure 5.2). In order to design efficient remedies against ETEC, its multiplication and host effector mechanisms should be exploited. Proteins which perform various pathogenic activities have been investigated in this study. One of the most striking features of ETEC strains is the presence of high percentage of identified insertion (IS) elements and mobile features, comprising approximately 4% of the genome. Existence of IS elements may provide a mechanism by which the plasmids can integrate into or expunge from the chromosome [34,35]. ETEC chromosome contains a unique gene cluster, EcE24377A_2278 to EcE24377A_229 responsible for the propanediol utilization, which was lost in an ancestor of *E. coli* and *Salmonella*, and reintroduced as a result of a relatively recent horizontal gene transfer event leading to altered metabolism. In Figure 5.2 generated using Civi [36], it can be observed that no genome-wide strand bias was observed. Another interesting aspect of the E24377A genome is the presence of multiple plasmids as compared to other pathogenic *E. coli* isolates. Each plasmid contains virulence factors recognized as essential in ETEC pathogenesis. ETEC E24377A genome is undergoing swift change through gene rearrangements and appears as a pathovar in genetic flux [34]. Complete annotated genome and its features facilitate understanding of factors which mediate the immune

responses.

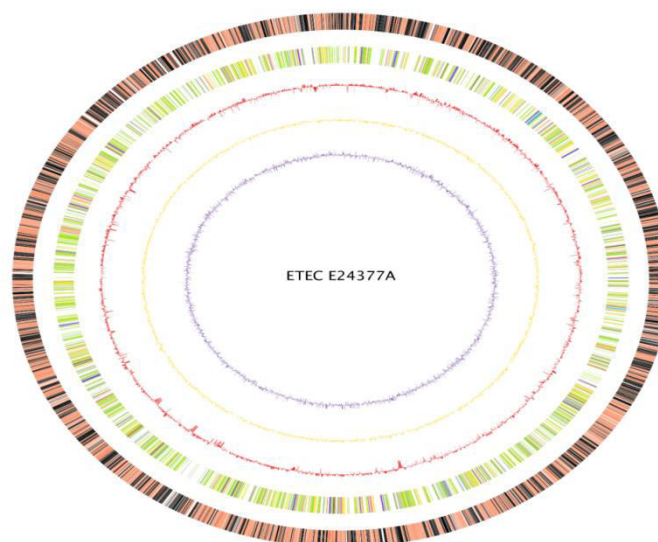


Figure 5.2: Circular genome of ETEC E24377A.

(i) ORFs plus strand: The outer most ring in black color shows ORFs plus strand. In the same ring orange regions indicate the ORFs negative strand (ii) Subcellular localization: In the adjacent ring, green, yellow, purple, light blue, orange, red, and gray color represents cytoplasm, cytoplasmic membrane, outer membrane, extracellular region, cell wall, periplasmic region and unknown subcellular localization respectively (iii) The third red ring represents GC percentage. Peaks represent an increase in GC percentage (iv) fourth ring in yellow color shows GC skew values for the whole genome. (v) The innermost purple colored ring shows AT skew values calculated for the whole genome. The figure is generated by Circular Visualization for microbial genomes (CiVi).

5.3.2 Comparative genomics analysis

We sought to screen the entire protein repertoire of ETEC strains E24377A and H10407 to identify sequences with high probability of being protective antigens. Comparative genomics of pathogenic and commensal strains can shed light on disease-related genes and vaccine candidates. Thus, we identified proteins shared between pathogenic ETEC strains which were lacking in commensal *E. coli*. First, the shared protein pool between E24377A and H10407 ETEC strains was determined by a reciprocal BLAST analysis. Next, proteins present in the

commensal *E. coli* strain SE11 genomes were eliminated from this shared pool. Protein sequences from this shared pool were filtered based on length criterion. We subjected these protein sequences to a homology search against Human proteome database using BLASTP program from standalone blast suite and retrieved pathogen proteins which were non-homologous to humans. Identifying proteins non-homologous to humans is essential as it excludes the possibility of the peptide vaccine targeting hosts enzymes, thus avoiding adverse effects on humans [37]. This analysis rendered 61 protein sequences.

5.3.3 Antigenicity and transmembrane prediction

The VaxiJen server used to assess the antigenicity of the protein sequences predicted 9 proteins as antigenic above a threshold of 0.6 which were further scrutinized for their cellular location. It revealed that 3 proteins were membrane localized by combined predictions of PSORTb and TMHMM. After filtering based on physicochemical properties responsible for establishing a protein as a potential vaccine candidate, targeted proteins pooled out of surface proteome were further used for epitope mapping (Figure 5.3). The venn diagram representing 3 prioritized target proteins was prepared using Jvenn. [38]

5.3.4 T-cell epitope prediction

NetCTL predicted T-cell epitopes from each sequence against MHC-supertypes. 50 epitopes were predicted from the membrane localized antigenic sequences. These epitopes were further assessed by VaxiJen for antigenicity. A list of all the predicted epitopes is provided. As shown in Table 5.1, based on the combined informatics analysis, vaccine candidates were divided in three categories high priority (5 candidates), medium priority (11 candidates), and low priority (9 candidates). For each high priority epitope, SMM based IEDB MHC I processing prediction tool retrieved the MHC-I alleles with IC50 value less than 500 nM which were potential epitope binders. MHCpred predictions of MHC-I and MHC-II alleles as efficient epitope binders were taken together with IEDB tool predictions to generate a final list of potential binders for each high priority epitope. The results are summarized in Table 5.2 which lists potential MHC-binders along with their MHC-I and II alleles and their respective binding affinity values.

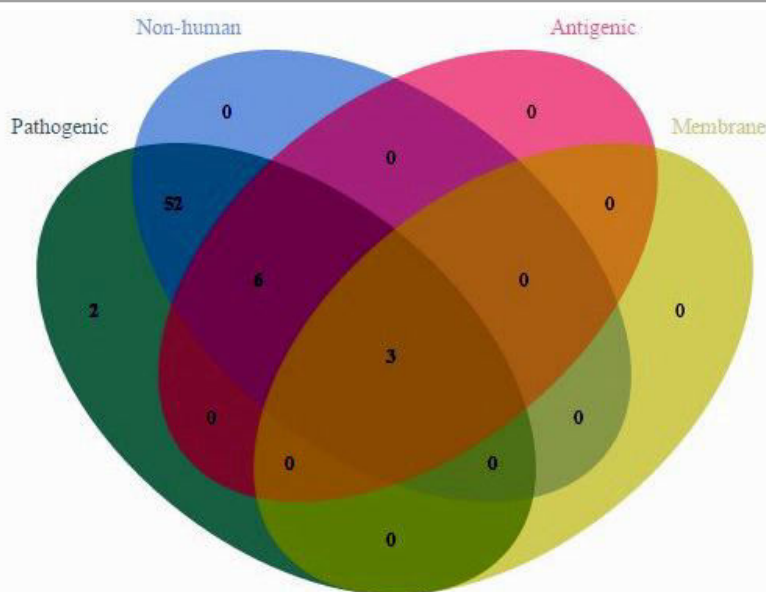


Figure 5.3: Representation of 3 prioritized target proteins: The proteins characterized as i) shared between pathogenic strains and excluded from commensal strain ii) Non-homologous to humans iii) Antigenic iv) Membrane localize, are shown using different colors in the Venn diagram. Proteins satisfying a particular parameter are shown in the corresponding category of the Venn diagram. Three proteins were prioritized for vaccine candidate identification. The image has been generated by Jvenn.

5.3.5 HLA-distribution analysis

For population coverage analysis, predicted high priority epitopes were clustered into two epitope sets based on the protein sequence from which they were generated. Population coverage analysis was then performed with these two epitopic sets along with their associated MHC-I and MHC-II alleles predicted through IEDB tool and MHCpred as input to IEDB population coverage analysis tool. As shown in Table 5.3, immune response elicitation of the 97.58% and 95.57% total world population was covered by the two predicted epitopic sets. Maximum coverage 99.23% for epitope set I was in European region followed by 97.53%, 94.42%, and 93.44% in the population of North American area, South Asia and East Asia respectively. For epitope set II, maximum coverage 97.25% was in Europe area followed by North America, East Asia, and Southeast Asia with coverage 96.01%, 96.01%, and 93.05% respectively.

Table 5.1: Most probable predicted epitopes selected on the basis of their NetCTL (MHC binding, proteasomal processing and TAP transport) and VaxiJen (Antigenicity) Score. Epitopes are categorized into High, Medium and Low priority epitopes based on antigenicity scale (VaxiJen Score).

Antigenicity Scale:**0.7-1.0:** Low Priority **1.0-1.7:** Medium Priority **1.7 and above:** High Priority

Priority	Epitope Set	Uniprot Id	Protein Name	Epitope	NetCTL Score	VaxiJen Score
High	Set I	A7ZGR5	Putative membrane protein	LICFFTLSY	1.8004	1.7024
				PLNPLILLY	1.7020	1.7194
				PIVNLFLLY	1.0515	1.9812
Medium	Set II	A7ZGK4	Uncharacterized protein	SVSVFIFLF	0.9647	3.0057
				SVFIFLFIY	1.0519	4.1539
				IIAFYEFMY	1.1235	1.6211
	Set I	A7ZGR5	Putative membrane protein	FYEFMYINY	1.1529	1.1573
				SLFGPEFLY	0.7607	1.5082
				KTALLICFF	0.9668	1.2435
				FTLSYNVLY	3.2100	1.1549
				STTIHSLFF	1.1339	1.2332
				KASNAHQRY	1.4873	1.2538
		Set II	O-antigen polymerase	ASHATTAGY	1.5819	1.1027
				KTTLYTINF	0.8133	1.1789
				YTINFMLSL	0.9437	1.6805
				SVGARLAMY	0.9550	1.0684
				TLLLGVLIIY	1.0991	0.7237
				TLSYNVLYF	0.8201	0.9366
Low	Set I	A7ZGR5	Putative membrane protein	VSTTIHSLF	1.1339	0.7892
				YYRFNDLFY	0.8111	0.7745
				YSSTKNIHQ	0.9350	0.8034
				MVRACIQMY	0.8529	0.8009
	Set II	A7ZGK4	Uncharacterized protein	SLATQLLFF	0.9629	0.7659
				TTAGYIILF	1.6656	0.9246
				FSAILIYAL	0.7571	0.7841

5.3.6 Molecular docking studies of HLA-epitope

Using Hex, different conformations of the predicted epitopes were generated and the best conformation was selected based on binding affinity scores. The docked complexes prepared in

Pymol and the interactions involved in HLA-A*11:01 binding with predicted epitopes are shown in Figure 5.4-5.8. HLA-A*11:01 binds with epitopes LICFFTLSY, PLNPLILLY, PIVNLFLLY, SVSVFIFLF, and SVFIFLFIY with docking energies -405.02, -404.60, -408.70, -400.57 and -446.00 respectively. Epitopes LICFFTLSY, PLNPLILLY, PIVNLFLLY interact with HLA-A*11:01 through interactions with MHC residues Arg 6, Asp 29, Asp 30, Pro 210, Glu 212. MHC interacts with epitopes SVSVFIFLF and SVFIFLFIY through hydrophobic interactions with Ser 4, Phe 8, Tyr 27, Asp 29, Asp 30, Met 98, Ala 211, and Glu 212. HLA-A*11:01 residues Arg 6 and Thr 233 are hydrogen bonded to Ser 1 and Phe 7 in epitope SVFIFLFIY. Involvement of common residues in interaction with different peptides suggests the crucial role of Arg 6, Asp 29, Asp 30, and Glu 212 MHC residues in MHC-peptide binding.

5.3.7 B-cell epitope identification

As per the criteria set for prediction of B-cell epitopes, the sequential epitopes were predicted using combined predictions of BCPREDS server and IEDB tools; further filtered based on antigenicity values from VaxiJen. Antigenic regions common to both BCPred and AAP algorithms at BCPREDS server were subjected to IEDB Emini surface accessibility tool to predict peptides which were surface exposed. Predicted peptides were checked for flexibility and hydrophilicity using IEDB tools Karplus & Schulz flexibility prediction, and Parker hydrophilicity prediction.

To test these peptides for their potential as B-cell epitopes, they were checked for their antigenicity which yielded five sequential B-cell epitopes with antigenicity score >0.09 (Table 5.4). Antibody interacting residues in membrane localized proteins predicted using CBTOPE server are summarized in Table 5 along with their secondary structure conformation. The probability of finding Conformational B cell epitopes is more in the coiled or turn region than in the helix or beta sheet [31]. Majority of the antibody interacting residues from membrane proteins were adopting coil conformation validating the accuracy of the predicted residues being a part of the conformational B-cell epitope. Models predicted from I-TASSER were selected based on their C-scores (confidence score) which is a measure of the quality of predicted models and based on the significance of threading template alignments and the convergence parameters. Residues which are predicted to be a part of conformational B-cell epitopes by DiscoTope server

are listed in Table 5.5. Only the residues with their DiscoTope score greater than 1 were considered.

Table 5.2: Predicted potential T cell epitopes, along with their interacting MHC-I and MHC-II alleles with an affinity of <500 nM and corresponding IC50 values (bracketed)

Epitope	#MHC-peptide binders	MHC-I alleles	MHC-II alleles
LICFFTLSY	24	HLA-A*01:01 (92.47), HLA-A*02:06 (105.68), HLA-A*02:17 (130.82), HLA-A*03:01 (305.49), HLA-A*11:01 (257.22), HLA-A*29:02 (30.89), HLA-A*30:02 (400.37), HLA-A*32:01 (468.75), HLA-A*32:07 (23.72), HLA-A*32:15 (164.53), HLA-A*68:01 (59.57), HLA-A*68:02 (492.04), HLA-A*68:23 (99.25), HLA-A*80:01 (26.42), HLA-B*15:01 (54.83), HLA-B*15:02 (293.24), HLA-B*15:03 (56.51), HLA-B*15:17 (212.18), HLA-B*27:20 (14.40), HLA-B*35:01 (133.53), HLA-B*40:13 (179.03), HLA-C*12:03 (14.73), HLA-C*14:02 (468.40)	HLA-DRB1*01:01 (9.40)
		HLA-A*01:01 (285.76), HLA-A*02:03 (4.80), HLA-A*02:17 (280.32), HLA-A*03:01 (323.59), HLA-A*11:01 (144.88), HLA-A*29:02 (25.63), HLA-A*31:01 (210.38), HLA-A*32:07 (20.42), HLA-A*32:15 (47.78), HLA-A*68:01 (127.64), HLA-A*68:23 (91.57), HLA-A*80:01 (2.30), HLA-B*15:02 (144.62), HLA-B*27:20 (4.76), HLA-B*35:01 (436.52), HLA-B*40:13 (399.88), HLA-C*05:01 (102.02), HLA-C*07:01 (326.72), HLA-C*07:02 (295.24), HLA-C*12:03 (48.11), HLA-C*14:02 (228.36)	HLA-DRB1*01:01 (1.01), HLA-DRB1*04:01 (359.75)
		HLA-A*01:01 (14.13), HLA-A*02:01 (456.04), HLA-A*02:03 (280.54), HLA-A*02:06 (38.55), HLA-A*03:01 (20.94), HLA-A*11:01 (49.32), HLA-A*26:01 (450.41), HLA-A*26:02 (108.22), HLA-A*29:02 (28.96), HLA-A*32:07 (26.92), HLA-A*32:15 (52.63), HLA-A*68:01 (99.54), HLA-A*68:23 (76.34), HLA-A*80:01 (4.53), HLA-B*15:02 (159.31), HLA-B*27:20 (35.69), HLA-B*35:01 (251.77), HLA-B*40:13 (93.53), HLA-C*03:03 (255.63), HLA-C*12:03 (23.62)	HLA-DRB1*01:01 (0.78), HLA-DRB1*07:01 (47.64)
		HLA-A*02:01 (306.90), HLA-A*02:02 (269.78), HLA-A*02:06 (307.61), HLA-A*02:50 (74.00), HLA-A*11:01 (6.90), HLA-A*24:02 (467.34), HLA-A*24:03 (388.63), HLA-A*26:02 (63.58), HLA-A*29:02 (325.66), HLA-A*32:01 (111.41), HLA-A*32:07 (7.91), HLA-A*32:15 (179.99), HLA-A*68:01 (50.23), HLA-A*68:02 (202.77), HLA-A*68:23 (11.72), HLA-B*15:02 (222.45), HLA-B*15:03 (495.55), HLA-B*15:17 (93.69), HLA-B*27:20 (70.39), HLA-B*35:01 (242.66), HLA-B*40:13 (106.40), HLA-B*58:01 (454.26), HLA-C*03:03 (92.17), HLA-C*07:01 (239.98), HLA-C*07:02 (297.97), HLA-C*12:03 (84.96)	HLA-DRB1*01:01 (12.45)
PIVNLFLLY	22	HLA-A*02:01 (316.96), HLA-A*02:02 (467.74), HLA-A*02:06 (77.62), HLA-A*02:17 (90.50), HLA-A*03:01 (239.33), HLA-A*11:01 (18.84), HLA-A*26:01 (454.58), HLA-A*26:02 (73.17), HLA-A*29:02 (21.27), HLA-A*30:02 (77.89), HLA-A*32:01 (253.48), HLA-A*32:07 (13.09), HLA-A*32:15 (53.12), HLA-A*68:01 (15.56), HLA-A*68:02 (283.79), HLA-A*68:23 (5.53), HLA-A*80:01 (18.37), HLA-B*15:02 (157.12), HLA-B*15:17 (46.96), HLA-B*27:20 (41.83), HLA-B*35:01 (160.91), HLA-B*40:13 (20.37), HLA-C*07:02 (444.81), HLA-C*12:03 (11.07)	HLA-DRB1*01:01 (41.59)
SVSVFIFLF	27		
SVFIFLFYI	25		

Table 5.3: Population coverage of predicted epitopes (Set I and Set II) based on MHC-I and MHC-II restriction data. **5.3a)** Maximum population coverage by South Africa **5.3b)** Maximum population coverage by Europe. Set I represents epitopes LICFFTLSY, PLNPLILLY, and PIVNLFLLY from protein with Uniprot Id A7ZGR5. Set II represents epitopes SVSVFIFLF and SVFIFLFIY from protein with Uniprot Id A7ZGK4

Population / Area	Class I and II Coverage	Population / Area	Class I and II Coverage
World	97.58%	World	95.57%
East Asia	93.44%	East Asia	96.01%
Northeast Asia	92.26%	Northeast Asia	91.23%
South Asia	94.42%	South Asia	90.26%
Southeast Asia	87.54%	Southeast Asia	93.05%
Southwest Asia	89.71%	Southwest Asia	85.79%
Europe	99.23%	Europe	97.25%
East Africa	86.70%	East Africa	85.47%
West Africa	88.20%	West Africa	88.68%
Central Africa	84.79%	Central Africa	84.55%
North Africa	93.27%	North Africa	88.91%
South Africa	91.71%	South Africa	91.08%
West Indies	92.42%	West Indies	90.89%
North America	97.53%	North America	96.01%
Central America	14.41%	Central America	4.85%
South America	89.09%	South America	87.93%
Oceania	80.85%	Oceania	95.57%

5.3a

5.3b

5.4 Discussion

The advent of high throughput sequencing techniques has enabled researchers to move beyond the classical vaccinology approach and employ computers for rational vaccine development with the information present in the genome. Vaccine development against the meningitis marks the beginning of reverse vaccinology (RV) technique [39]. Hitherto, the use of omics data together with bioinformatics approaches has been the hallmark of vaccinology in the genomics era. Since then, RV has been practically applied against many pathogens [40,41]. Omics-guided approaches have been applied to its full potential by identifying vaccines against several human and animal pathogens such as *Staphylococcus aureus*, *Salmonella* spp., *Neisseria meningitides* (Serogroup B), *Bacillus anthracis*, *Bordetella pertussis*, West Nile Virus, *Candida albicans*, *Plasmodium*

falciiparum, *Streptococcus canis* etc., now being in late stage clinical trials or preclinical studies [42].

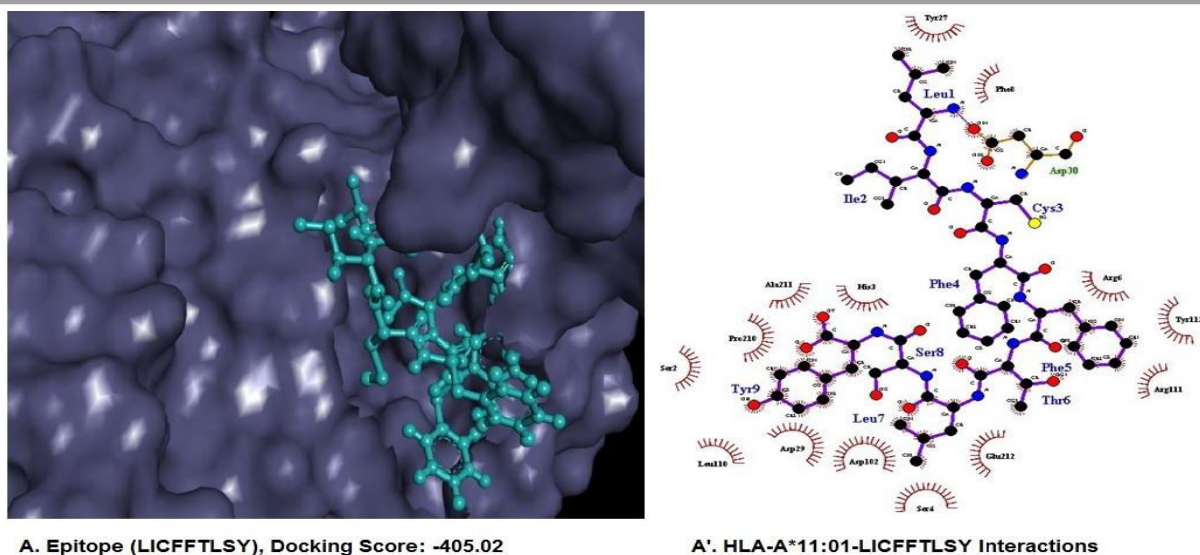


Figure 5.4: A) Docked complex of HLA-A*11:01 with epitope LICFFFTLSY visualized in Pymol A') and corresponding interactions involved in binding visualized in LigPlot.

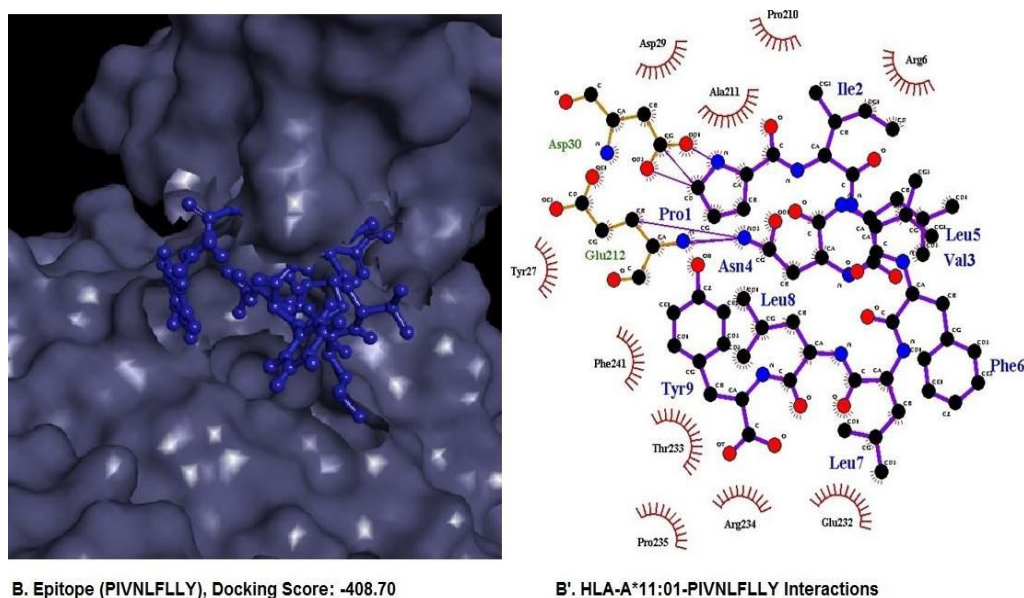
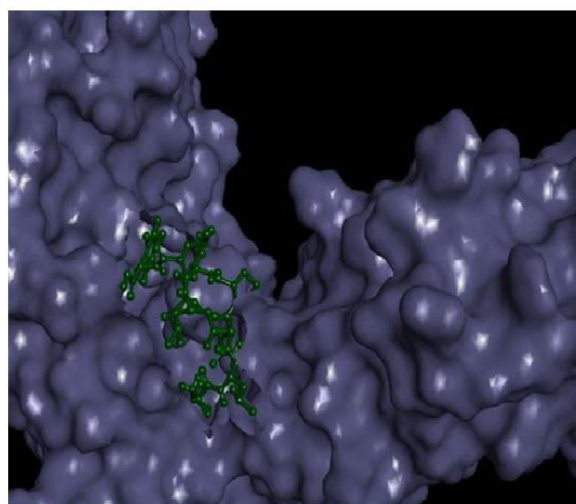


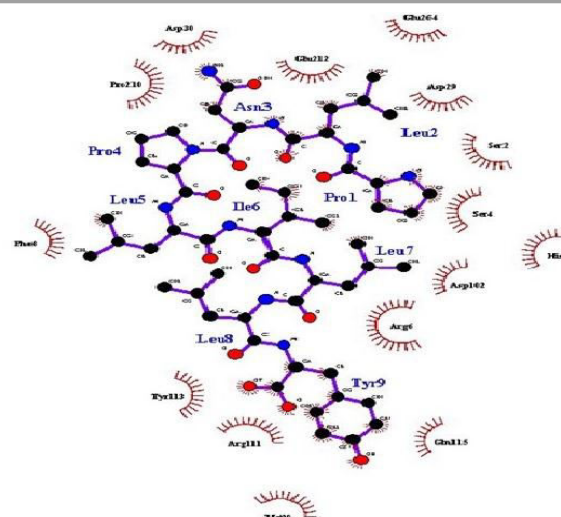
Figure 5.5: B) Docked complex of HLA-A*11:01 with epitope PIVNLFLLY visualized in Pymol B') and corresponding interactions involved in binding visualized in LigPlot.

The design of an epitope-based vaccine candidate against ETEC has already been proposed but these studies are focused on some specific proteins like autotransporter protein, adhesion

proteins like eae (intimin), colonization factors like CFA/1, CS2 and CS3 and invasion proteins like invasion plasmid antigen C (IpaC), chimeric multi subunit proteins [43-46]. Subunit ETEC vaccine research has been mainly focused on well known and characterized virulence factors such as toxins, major colonization factors, and adhesion proteins [7] which are essential for the inception of ETEC colonization and/or host damage. Yet, the genome of ETEC strains may encode several unveiled antigenic proteins, which have so far not been explored as vaccine candidates.

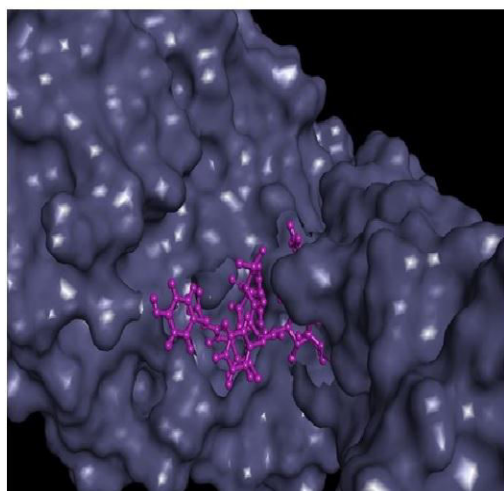


C. Epitope (PLNPLILLY), Docking Score: -404.60

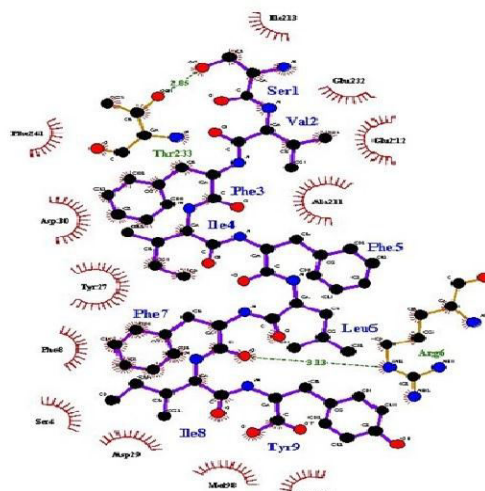


C' HLA-A*11:01-PLNPLILLY Interactions

Figure 5.6: C) Docked complex of HLA-A*11:01 with epitope PLNPLILLY visualized in Pymol C') and corresponding interactions involved in binding visualized in LigPlot.



D. Epitope (SVFIFLFIY), Docking Score: -446.00



D'. HLA-A*11:01-SVFIFLFIY Interactions

Figure 5.7: D) Docked complex of HLA-A*11:01 with epitope SVFIFLFIY visualized in Pymol **D')** and corresponding interactions involved in binding visualized in LigPlot.

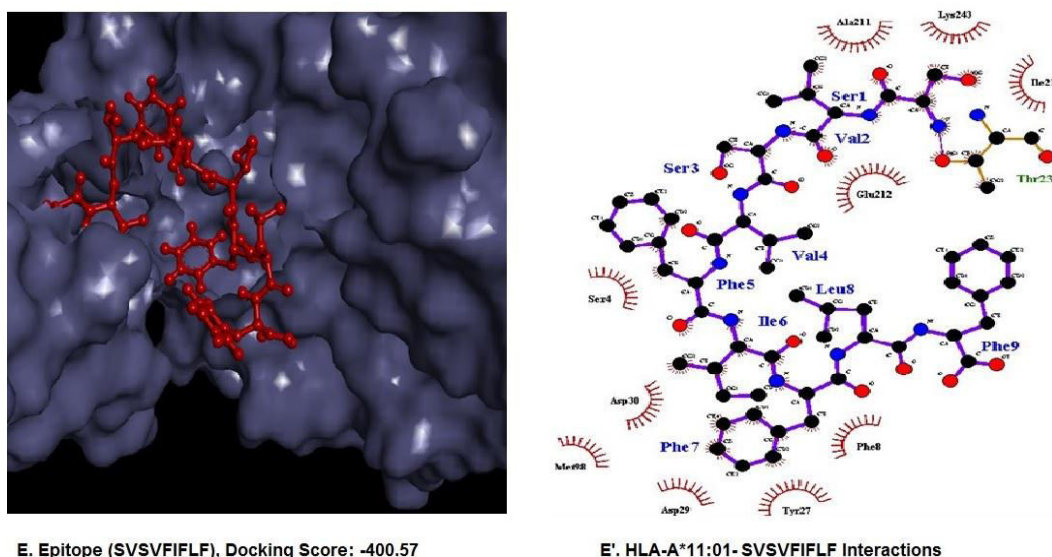


Figure 5.8: E) Docked complex of HLA-A*11:01 with epitope SVSVFIFLF visualized in Pymol **E')** and corresponding interactions involved in binding visualized in LigPlot.

Omics techniques have been projected as a powerful aid in vaccine development, particularly for the pathogens for which pathogenesis mechanism or antigenic determinants knowledge is inadequate. In this study, we have used integrated comparative genomics and immunoinformatics analysis of available ETEC genomes in the search for vaccine candidates. We have used an unbiased screening approach by examining all encoded sequences in ETEC genome, without considering their annotated function, which facilitates access to obscured potential candidates overlooked by previous ETEC vaccine studies. Similar studies have been performed by a group of researchers to identify potential vaccine candidate against Enterohemorrhagic *Escherichia coli* (EHEC) O157:H7, a highly virulent serotype implicated in outbreaks and sporadic cases of diarrhea which infects colon by producing relatively large amounts of the bacteriophage-mediated Shiga-like toxin [47,48]. Presently vaccines are based on antibody mediated immunity (AMI) provided by B-cells. Antibodies are currently the only way that confer immediate immunological protection against biological weapons in populations which are immunologically immature [49]. However, T-cells confer long-lasting immunity while antibody mediated immunity can be easily overcome by antigenic drift [50]. In this study, we

have proposed both B and T-cell epitopes which could be tested through *in vivo* and *in vitro* experiments for their efficacy in eliciting humoral and cell mediated immunity.

In order to accelerate the process of identifying highly antigenic vaccine candidates using immunoinformatics approach, we have worked around a set of criteria that a predicted peptide must meet in order to qualify as potential vaccine candidate. As depicted in Figure 5.1, the proteins were selected based on physicochemical properties like surface localization (secreted), or having transmembrane (TM) domains. The VaxiJen server predicted the protective antigens based on the overall antigenicity score and assisted to prioritize the proteins. While screening genomic data, it is of utmost importance to choose the genomic portions with the most probable protective antigens. Such genomic segments can be identified by searching for sequences with antigenic potential. Antigenicity filter was employed at several stages; first to identify the antigenic proteins and at a later stage to identify peptides with high antigenicity values.

For T-cell epitope predictions, parameters such as T-cell processivity, the number of human leukocyte antigens (HLA) alleles covered, and significant population coverage. The characteristics features of B-cell epitopes include physicochemical propensities like flexibility, hydrophilicity, solvent accessibility which have been utilized in many B-cell prediction programs [51]. For sequential B-cell epitope predictions, we identified epitopes which could be efficiently processed by B-lymphocytes and were chosen based on criteria of surface accessibility, flexibility and hydrophilicity. Majority of the B-cell epitopes (~90%) are conformational epitopes. CBTOPE which predicts conformational B-cell epitopes in an antigen from its amino acid sequence was used as for the proteins identified as vaccine candidates, 3D structure was not available. Subsequently, this approach rendered a list of five high priority T-cell epitopes, five sequential B-cell epitopes and antibody interacting residues forming conformational B-cell epitopes as potential vaccine candidates. B-cell epitopes most likely to induce immune responses were from O-antigen polymerase, putative membrane protein and an uncharacterized protein. Potential T-cell epitopes were chosen by using NetCTL server results based on integrated predictions of peptide MHC class I binding, proteasomal C terminal cleavage and TAP transport efficiency. Predicted peptides showed good population coverage despite the fact that in case of MHC-II data was only available for the alleles HLA-DRB1*01:01, HLA-DRB1*04:01, and HLA-DRB1*07:01. However as is clear from the Table 5.3, all the predicted nonamers were

interacting with most common HLA allele HLA-DRB1*01:01. As shown in Table 5.4, population coverage in developing world was highest in Asia and Africa where the diarrheal incidence rate is highest. In industrialized nations population coverage was highest for Europe and North America as these countries have maximum number of travelers to Asian and African regions [43].

Table 5.4: Five most potential B-cell epitopes by combined predictions of AAP, BCPred, IEDB tools (Emini Surface Accessibility, Karplus and Schulz flexibility, Parker hydrophilicity), further filtered based on their antigenicity values (VaxiJen Score).

Uniprot Id	Start Residue	End Residue	Epitope	Length	VaxiJen Score
A7ZGR5	345	352	PETHKSDN	8	1.1952
	237	246	FKNQFNKKIT	10	0.9161
	179	189	YSSTKNIHQK	11	1.0160
A7ZTH5	272	279	YSHDNTRT	8	1.9700
	303	319	EKRAEKIHELEEKEPRL	17	1.0362

Further investigation in the data shows that epitopic set II of high priority T-cell epitopes has high antigenicity value (Table 5.2) and has maximum number of MHC-interacting alleles. SVSVFIFLF has highest number of interactors, showing interaction with a total of 27 MHC-I and MHC-II alleles. Epitope sequence SVFIFLFIY from epitopic set II has strongest binding affinity (316.96 nM) for MHC supertype (HLA-A*02:01). The results from population coverage analysis reveal that epitopic set I and II both cover different populations effectively. Five potential B-cell epitopes were predicted with their VaxiJen score (antigenicity) greater than 0.9. The epitopes from set II, SVSVFIFLF and SVFIFLFIY have highest antigenicity of 3.0057 and 4.1539 respectively. Elicitation of effective immune responses depends on the specificity and diversity of the T-cell epitopes binding to HLA-alleles. Besides due to highly polymorphic nature of MHC, identifying peptides which can bind to as many MHC-alleles is a requisite [52]. Findings from the current study show that all the predicted epitopes bound to more than 20 MHC-alleles and had extensive human population coverage.

All the predicted T cell epitopes were found to interact with HLA-A*11:01 with varying affinities. Most probable immunogenic T-cell epitopes were from membrane proteins and some

uncharacterized proteins. Thus, HLA-A*11:01 was selected for docking studies. The results of computational docking study show that all the peptides bind efficiently to the HLA alleles. In summary, our comparative genomic approach integrated with immunoinformatics analysis allowed us to identify a group of ETEC specific vaccine candidates, some of them with high potential to encode for protective immunogens. This approach confirms the antigenic and protective efficacy of a subset of the candidates and further requires the experimental validation of the predicted epitopes in order to increase the therapeutic arsenal of vaccines against ETEC. Moreover, uncharacterized proteins may be associated with virulence in ETEC, and their characterization could offer new insights into ETEC pathogenesis. Consequently, the predicted vaccine targets open up new avenues for the study of functional characterization of these proteins regarding interaction with host and vaccine development.

Table 5.5: Potential conformational B-cell epitope residues predicted using CBTOPE and DiscoTope server, along with the corresponding secondary structure conformation each residue adopts and DiscoTope scores.

Uniprot Id	CBTOPE					DiscoTope		
	Start Residue	End Residue	Epitope	Length	Secondary Structure Type	Residue No	Residue	DiscoTope Score
A7ZGR5	322	322	I	1	C	238	K	2.081
	342	345	VLFP	4	CCCC	325	Q	1.718
	349	351	KSD	3	CCC	327	I	1.920
	404	405	IR	2	EC	328	N	3.520
						363	L	2.118
						366	K	1.393
A7ZGK4	17	17	V	1	C			
	136	139	PDKR	4	CCCC			
A7ZTH5	291	293	GLK	3	HHH			
	324	324	P	1	C			
	410	410	S	1	C			

There are certain limitations to our approach for vaccine identification in that the tools utilized in the computational pipeline have some inherent technical challenges. As these tools are data driven, the accuracy depends on the quality of the experimental data used to train these models. Even the experimental data have the potential to be inaccurate due to experimental errors or flawed interpretations. Standard docking tools are meant for protein-protein or protein-ligand docking where active sites are strongly conserved to preserve protein functions. While antigen-

antibody binding sites are less conserved due to the competition for survival against the host immune system, thus using these docking tools presents us with some limitations. The computational approach for vaccine identification is not a substitute for experimental approaches but a potent complementary approach to reduce time and effort invested in culture based experimental methods.

Conclusion

Vaccine identification using traditional molecular immunology techniques is a time-consuming, arduous and expensive task. With a myriad of omics techniques, availability of genomic sequence data and novel bioinformatics approaches, our understanding of complex human diseases has radically improved. These techniques have accelerated the pace of vaccine development as they hold the potential for rapid screening and identification of probable vaccine candidates from the whole genome; thus significantly reducing the epitope dataset to be tested experimentally. Our selection criteria based on integrated approach of comparative genomics and immunoinformatics aided in categorizing the candidates as high, medium, and low priority. The high priority epitopes predicted in our study have greater than 95% population coverage, thus being a representative of larger human population and interacts with major MHC supertypes. Our findings may help to develop vaccines against ETEC once validated experimentally using model organisms.

References

1. Gupta SA K, Keck J, Ram PA K, Crump JA A, Miller MA A, Mintz EA D. Part III. Analysis of data gaps pertaining to enterotoxigenic *Escherichia coli* infections in low and medium human development index countries, 1984-2005. *Epidemiology and Infection*, 136(6), 721-738 (2008).
2. Guerrant RL, Kosek M, Moore S, Lortz B, Brantley R, Lima AAM. Magnitude and Impact of Diarrheal Diseases. *Archives of Medical Research*, 33(4), 351-355 (2015).
3. Qadri F, Svennerholm AM, Faruque ASG, Sack RB. Enterotoxigenic *Escherichia coli* in Developing Countries: Epidemiology, Microbiology, Clinical Features, Treatment, and Prevention. *Clinical Microbiology Reviews*, 18(3), 465-483 (2005).
4. Del Canto F, Botkin DJ, Valenzuela P *et al.* Identification of Coli Surface Antigen 23, a Novel Adhesin of Enterotoxigenic *Escherichia coli*. *Infection and Immunity*, 80(8), 2791-2801 (2012).
5. Moriel DG, Scarselli M, Serino L, Mora M, Rappuoli R, Masignani V. Genome-based vaccine development: A short cut for the future. *Human Vaccines*, 4(3), 184-188 (2008).

6. Rinaudo CD, Telford JL, Rappuoli R, Seib KL. Vaccinology in the genome era. *The Journal of Clinical Investigation*, 119(9), 2515-2525 (2009).
7. Walker RI. An assessment of enterotoxigenic *Escherichia coli* and *Shigella* vaccine candidates for infants and children. *Vaccine*, 33(8), 954-965 (2015).
8. Fleischmann RD, Adams MD, White O *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223), 496-512 (1995).
9. Pizza M, Scarlato V, Masignani V *et al.* Identification of Vaccine Candidates Against Serogroup B *Meningococcus* by Whole-Genome Sequencing. *Science*, 287(5459), 1816-1820 (2000).
10. Moriel DG, Bertoldi I, Spagnuolo A *et al.* Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 107(20), 9072-9077 (2010).
11. Lawley TD, Croucher NJ, Yu L *et al.* Proteomic and Genomic Characterization of Highly Infectious *Clostridium difficile* 630 Spores. *Journal of Bacteriology*, 191(17), 5377-5386 (2009).
12. McCarthy A, Lindsay J. Genetic variation in *Staphylococcus aureus* surface and immune evasion genes is lineage associated: implications for vaccine design and host-pathogen interactions. *BMC Microbiology*, 10(1), 173 (2010).
13. Brocchieri L, Karlin S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, 33(10), 3390-3400 (2005).
14. Doytchinova I, Flower D. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, 8(1), 4 (2007).
15. Yu NY, Wagner JR, Laird MR *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26(13), 1608-1615 (2010).
16. Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics*, 8, 424-424 (2007).
17. Tenzer S, Peters B, Bulik S *et al.* Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cellular and Molecular Life Sciences CMLS*, 62(9), 1025-1037 (2005).
18. Peters B, Sette A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics*, 6, 132-132 (2005).
19. Guan P, Doytchinova IA, Zygouri C, Flower DR. MHCpred: a server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Research*, 31(13), 3621-3624 (2003).
20. Bui HH, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics*, 7, 153-153 (2006).
21. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Research*, 39(suppl 1), D913-D919 (2011).

22. Thevenet P, Shen Y, Maupetit J, Guyon Fdr, Derreumaux P, Tuffery P. PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic Acids Research*, 40(Web Server issue), W288-W293).
23. Macindoe G, Mavridis L, Venkatraman V, Devignes M-D, Ritchie DW. HexServer: an FFT-based protein docking server powered by graphics processors. *Nucleic Acids Research*, 38(Web Server issue), W445-W449 (2010).
24. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.3r1. (Ed.^(Eds) (2010)
25. Laskowski RA, Swindells MB. LigPlot+: Multiple Ligand-Protein Interaction Diagrams for Drug Discovery. *Journal of Chemical Information and Modeling*, 51(10), 2778-2786 (2011).
26. El-Manzalawy Y, Dobbs D, Honavar V. Predicting linear B-cell epitopes using string kernels. *Journal of molecular recognition : JMR*, 21(4), 243-255 (2008).
27. Chen J, Liu H, Yang J, Chou KC. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids*, 33(3), 423-428 (2007).
28. Emini EA, Hughes JV, Perlow DS, Boger J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *Journal of Virology*, 55(3), 836-839 (1985).
29. Karplus PA, Schulz GE. Prediction of chain flexibility in proteins. *Naturwissenschaften*, 72(4), 212-213 (1985).
30. Parker JMR, Guo D, Hodges RS. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites. *Biochemistry*, 25(19), 5425-5432 (1986).
31. Ansari HR, Raghava GPS. Identification of conformational B-cell Epitopes in an antigen from its primary sequence. *Immunome Research*, 6, 6-6 (2010).
32. Kringelum JV, Lundegaard C, Lund O, Nielsen M. Reliable B Cell Epitope Predictions: Impacts of Method Development and Improved Benchmarking. *PLoS Comput Biol*, 8(12), e1002829 (2012).
33. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols*, 5(4), 725-738 (2010).
34. Rasko DA, Rosovitz MJ, Myers GSA *et al.* The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates. *Journal of Bacteriology*, 190(20), 6881-6893 (2008).
35. Heap JT, Ehsaan M, Cooksley CM *et al.* Integration of DNA into bacterial chromosomes from plasmids without a counter-selection marker. *Nucleic Acids Research*, 40(8), e59-e59 (2012).
36. Overmars L, van Hijum SAFT, Siezen RJ, Francke C. CiVi: circular genome visualization with unique features to analyze sequence elements. *Bioinformatics*, (2015).
37. Butt AM, Nasrullah I, Tahir S, Tong Y. Comparative Genomics Analysis of Mycobacterium ulcerans for the Identification of Putative Essential Genes and Therapeutic Candidates. *PLoS ONE*, 7(8), e43080 (2012).
38. Bardou P, Mariette J, Escudie F, Djemiel C, Klopp C. jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics*, 15(1), 293 (2014).

39. Giuliani MM, Adu-Bobie J, Comanducci M *et al.* A universal vaccine for serogroup B meningococcus. *Proceedings of the National Academy of Sciences of the United States of America*, 103(29), 10834-10839 (2006).
40. Maione D, Margarit I, Rinaudo CD *et al.* Identification of a Universal Group B Streptococcus Vaccine by Multiple Genome Screen. *Science (New York, N.Y.)*, 309(5731), 148-150 (2005).
41. Thorpe C, Edwards L, Snelgrove R *et al.* Discovery of a vaccine antigen that protects mice from Chlamydia pneumoniae infection. *Vaccine*, 25(12), 2252-2260 (2007).
42. Bagnoli F, Baudner B, Mishra RPN *et al.* Designing the Next Generation of Vaccines for Global Public Health. *OMICS: A Journal of Integrative Biology*, 15(9), 545-566 (2011).
43. Harris JA, Roy K, Woo-Rasberry V *et al.* Directed Evaluation of Enterotoxigenic Escherichia coli Autotransporter Proteins as Putative Vaccine Candidates. *PLoS Neglected Tropical Diseases*, 5(12), e1428 (2011).
44. Bagheri S, Mousavi Gargari SL, Rasooli I, Nazarian S, Alerasol M. A CsaA, CsaB and LTB chimeric protein induces protection against Enterotoxigenic Escherichia coli. *The Brazilian Journal of Infectious Diseases*, 18(3), 308-314 (2014).
45. Nazarian S, Mousavi Gargari SL, Rasooli I, Amani J, Bagheri S, Alerasool M. An in silico chimeric multi subunit vaccine targeting virulence factors of enterotoxigenic Escherichia coli (ETEC) with its bacterial inbuilt adjuvant. *Journal of Microbiological Methods*, 90(1), 36-45 (2012).
46. Hashish EA, Zhang C, Ruan X *et al.* A Multiepitope Fusion Antigen Elicits Neutralizing Antibodies against Enterotoxigenic Escherichia coli and Homologous Bovine Viral Diarrhea Virus In Vitro. *Clinical and Vaccine Immunology : CVI*, 20(7), 1076-1083 (2013).
47. Garcia-Angulo VA, Kalita A, Kalita M, Lozano L, Torres AG. Comparative Genomics and Immunoinformatics Approach for the Identification of Vaccine Candidates for Enterohemorrhagic Escherichia coli O157:H7. *Infection and Immunity*, 82(5), 2016-2026 (2014).
48. Kalita A, Kalita M, Torres AG. Exploiting the power of OMICS approaches to produce E. coli O157 vaccines. *Gut Microbes*, 5(6), 770-774 (2014).
49. Casadevall A, Pirofski LA. New Concepts in Antibody-Mediated Immunity. *Infection and Immunity*, 72(11), 6191-6196 (2004).
50. Bacchetta R, Gregori S, Roncarolo M-G. CD4+ regulatory T cells: Mechanisms of induction and effector function. *Autoimmunity Reviews*, 4(8), 491-496 (2005).
51. Li X, Yang HW, Chen H, Wu J, Liu Y, Wei JF. In Silico Prediction of T and B Cell Epitopes of Der f 25 in Dermatophagoides farinae. *International Journal of Genomics*, 2014, 10 (2014).
52. Germain RN. MHC-dependent antigen processing and peptide presentation: Providing ligands for T lymphocyte activation. *Cell*, 76(2), 287-299 (1994).

CHAPTER 6

**Novel Drug Targets for Food-Borne Pathogen
Campylobacter jejuni: An Integrated Subtractive Genomics
and Comparative Metabolic Pathway Study**

6.1 Introduction

Campylobacters are a major cause of foodborne diarrheal illness and result in high morbidity and mortality rate, and economic loss in every region of the world [1]. In developing countries, *Campylobacter* infections are frequent in children under age two, sometimes leading to death. In industrialized nations, they are most frequently identified cause of bacterial diarrhea in early adulthood [2]. According to a report released by Center for Disease Control and Prevention (CDC), there are 1.3 million incidences of campylobacteriosis and there is rapid escalation of antibiotic resistance in *Campylobacter* from 13% in 1997 to almost 25% in 2011 in United States [3]. Growing body of literature has documented that resistance to antibiotics like quinolones, macrolides, tetracyclines, chloramphenicol, cephalosporins, aminoglycosides is increasing rapidly in most parts of the world due to common and indiscriminate use of these agents [4-7]. *Campylobacters* are highly important from socioeconomic perspective, which strongly indicates a need for novel therapeutic targets with a high potential to improve quality of life and survival rates.

Since the publication of pathogenic bacterial genome sequences *Haemophilus influenza* [8] and *Mycoplasma genitalium* [9] in 1995, the number of completed genome sequence for various microbial species has increased rapidly. This data in the post genomic era has provided researchers with the possibility to fully exploit it for identification of novel therapeutic targets and opened up new avenues for genome wide application of comparative and subtractive genomics approaches for therapeutic intervention.

Subtractive genomics approach has been adverbently used by many researchers [10-13] in search of novel drug targets for different microbes. Genome sequences of several *Campylobacter* species have been published including *Campylobacter concisus*, *Campylobacter curvus*, *Campylobacter fetus*, *Campylobacter hominis* and six strains of *Campylobacter jejuni* (<http://gcid.jcvi.org/projects/msc/campylobacter/>). In this study we report the subtractive genomics approach integrated with comparative metabolic pathway analysis aimed at identifying novel therapeutic target proteins of *C. jejuni* pathogenic strain NCTC11168.

6.2 Methods

Different databases and tools as described in the workflow (Figure 6.1) were utilized for the identification of putative therapeutic targets against *C. jejuni* integrating subtractive genomics approach with genome wide comparative pathway analysis.

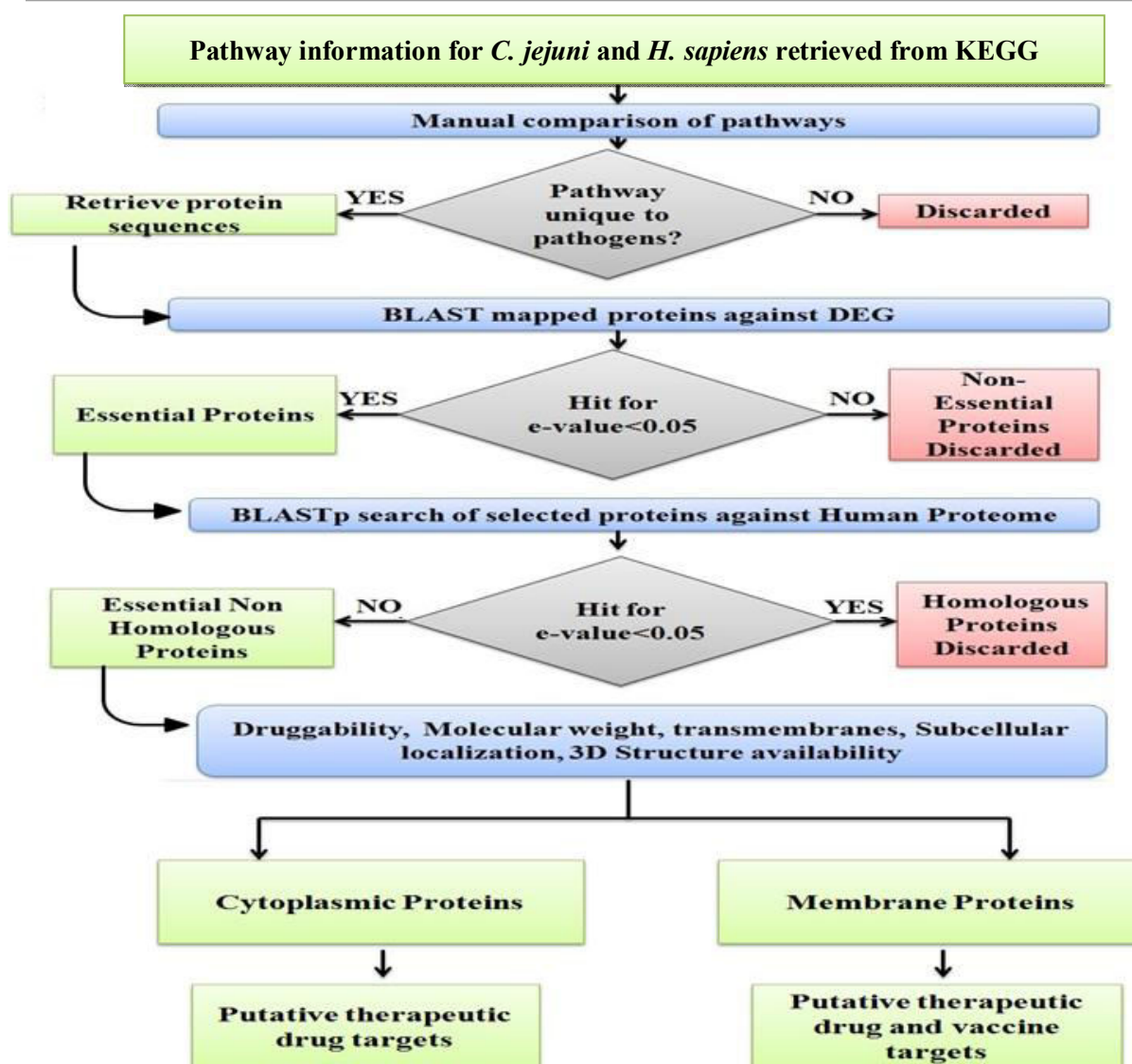


Figure 6.1: Schematic workflow of comparative genomics approach to identify drug targets in pathogenic *C. jejuni* strain NCTC11168.

6.2.1 Host and pathogen comparative metabolic pathway analysis

The Kyoto Encyclopedia of Genes and Genomes (KEGG) database [14], the most comprehensive resource of pathway information was used for comparative genome wide pathway analysis of the *C. jejuni* and *Homo sapiens*. Manual comparison was done to identify the pathways unique to *C. jejuni* as per KEGG database annotations. Protein sequences for the enzymes involved in the unique pathways were retrieved from the Uniprot protein database.

6.2.2 Subtractive genomics and identification of essential non-homologous pathogen proteins

Essential non-homologous proteins from pathogen proteome were selected by two step comparison. In first comparison, subtractive genomics approach was applied on pathogen proteins from unique pathways. The BLAST search [15] based on essentiality criteria was performed against Database of Essential Genes (DEG version 10.9) [16] which hosts records of essential genomic elements critical for an organism's survival, such as protein-coding genes and non-coding RNAs, among bacteria, archaea and eukaryotes. An expectation value (e-value) threshold of 0.05 was used as filtering criteria for BLAST hits; with *C. jejuni* and *Helicobacter pylori* as background organism against which similarity search for identification of essential genes was performed. In the second step comparison, essential pathogen proteins were further screened out on the basis of homology with the host proteome at an e-value cut off of 0.05. In BLASTP search, proteins that did not have any hits below e-value inclusion threshold were selected to be essential non-homologous proteins from *C. jejuni*.

6.2.3 Prioritization of essential non-homologous proteins for therapeutic targets

The molecular and structural properties of essential non-homologous pathogen proteins were calculated which aid in prioritization of the drug targets that will most likely lead to effective treatments. Subcellular localization of identified proteins was predicted using PSORTb [17] which uses a Bayesian network model to calculate associated probability value for five major localization sites viz. cytoplasmic, inner membrane, periplasmic, outer membrane and extracellular with p-value criteria 7.5. TMHMM server which is a Hidden Markov Model (HMM) based tool for prediction of alpha helices in membrane proteins was used for transmembrane predictions [18]. A search was performed to identify the proteins for which experimentally or computationally solved structures are available in PDB [19] or Modbase [20]

or ProteinModelPortal [21]. A search of identified proteins in Uniprot was performed for retrieving information pertaining to molecular weight and existence of proteins. The information on protein existence provides different types of evidence for experimental characterization at protein and/or transcript level, homology inference or uncertainty. VaxiJen server [22] was employed to check the antigenicity of the membrane localized proteins with a threshold of 0.6. The antigenic sequences were further screened for their ability to bind to MHC Class I molecule using ProPred-I [23]. ProPred-I implements proteasomal processing with matrices for 47 MHC Class-I alleles to identify the regions in the antigenic sequence which can act as potential MHC binders.

6.2.4 Druggability of essential non-homologous proteins

To evaluate the druggability potential that is ability of a particular biological target to bind with high affinity to known drugs of each identified therapeutic target; we subjected each protein to a BLASTP search against DrugBank with an e-value 0.01. DrugBank is a unique comprehensive resource that integrates drug data with drug target information at sequence, structure and pathway level [24]. DrugBank version 4.2 currently has information about 7737 drug entries including 1585 FDA-approved small molecule drugs, 158 FDA-approved biotech (protein/peptide) drugs, 89 nutraceuticals and over 6000 experimental drugs; along with 4281 non-redundant protein sequences (drug target/enzyme/transporter/carrier) linked to these drugs.

6.3 Results and discussion

6.3.1 Identification of unique metabolic pathways

The KEGG database initiated in 1995 is a computational representation for biological systems, integrating genetic information of genes and proteins with chemical and systemic information of molecular interaction and reaction networks. It links genetic building block information with higher order functional information. Currently KEGG database houses 87 different pathways for *C. jejuni* 11168 strain and 292 pathways for *H. sapiens*. As described in the workflow (Figure 6.1) manual comparison of host and pathogen pathways was conducted which resulted in identification of 16 unique pathways to the pathogen (Table 6.1) while 71 remaining pathways being shared by both humans and *Campylobacters*. Furthermore, the proteins involved in unique pathways were identified.

Table 6.1: A list of metabolic pathways unique to *C. jejuni* against human host and the number of proteins associated with the corresponding pathway

S. No	Pathway Name	Pathway ID	#Proteins
1	Carbapenem biosynthesis	cje0033	2
2	Novobiocin biosynthesis	cje00401	4
3	D-Alanine metabolism	cje00473	2
4	Streptomycin biosynthesis	cje00521	3
5	Lipopolysaccharide biosynthesis	cje00540	20
6	Peptidoglycan biosynthesis	cje00550	14
7	C5-Branched dibasic acid metabolism	cje00660	7
8	Methane metabolism	cje00680	15
9	Biosynthesis of secondary metabolites	cje01110	167
10	Microbial metabolism in diverse environments	cje01120	88
11	beta-Lactam resistance	cje01501	5
12	Vancomycin resistance	cje01502	5
13	Two-component system	cje02020	43
14	Bacterial chemotaxis	cje02030	20
15	Flagellar assembly	cje02040	33
16	Bacterial secretion system	cje03070	18

6.3.2 Identification of essential non-homologous pathogen proteins

A total of 446 proteins were identified to be involved in 16 unique pathways. Few proteins were involved in more than one pathway which resulted in 326 protein sequences from unique pathways after removing redundant protein sequences. A BLASTP search of these 326 protein sequences against 551 essential genes from *C. jejuni* and *H. pylori* in DEG was performed which revealed a total of 115 essential protein sequences with hits below e-value 0.05. Gene essentiality is thought to be important criteria for identification of therapeutic drug targets [25]. But there is a limitation to this approach as it fails to identify some targets such as hypoxanthine phosphoribosyl transferase as essential in *Plasmodium falciparum* (false negative) [26] while sometimes it yields false positives as in case of dihydrofolate reductase in *Leishmania major* [27]. Gene essentiality prediction via experimental methods such as single gene knockouts, RNA interference, and conditional knockouts is labor-intensive, expensive, and time-consuming [28]. Furthermore only few infectious agents are amenable to experimental approaches of gene essentiality as the tools for identification of drug or vaccine targets are often limited or absent for many pathogens. In such scenarios, computational methods for gene

essentiality prediction seem to streamline the gap between the amount of data generated from sequencing projects and whole genome approaches for prediction of essential genes [25,29,30].

A previous work also identified the drug targets in *C. jejuni* by utilizing the CAI (Codon Adaptation Index) criterion as a measure of gene essentiality [31]. Essential genes are highly conserved, highly expressed and preferentially positioned in the leading strand. But high gene expression rates do not significantly correlate with the gene strand biasness and non-essential genes also show high expression rates [32,33]; in which case the definition of essentiality based on high expression could be erroneous. In our study, we have utilized subtractive genomics approach to predict genes essential to *C. jejuni* through homology search against experimentally predicted essentiality data from *C. jejuni* and *H. pylori* in DEG, both of which belong to the epsilon class of proteobacteria. These essential protein sequences were further filtered out by homology search against human proteome for identification of non-homologous protein sequences. This comparison detected 66 essential non-homologous proteins with no hits against *H. sapiens* below e-value 0.05. This comparison was performed to identify proteins unique to pathogen so as to avoid adverse effects on the human host [28], as the potential drug may also target host's enzymes. These 66 essential non-homologous protein sequences represent the potential to be further exploited for therapeutic drug design against *C. jejuni*.

6.3.3 Prioritization of essential non-homologous proteins for therapeutic drug targets

Although all 66 essential non-homologous identified proteins are potential drug targets, yet these can be further filtered; using additional parameters that are determinant of structural and molecular properties of proteins such as molecular weight, subcellular localization, transmembrane prediction and availability of 3D structure; for prioritization of drug targets [25] which maximizes the likelihood of landing to the best therapeutic target against pathogen and reduces the time and resources for developing such an agent. Subcellular localization prediction using PSORTb program identified 63 proteins to be of cytoplasmic origin and all the essential non-homologous proteins were found to be <110 kDa; suggesting possibility of experimental verification of the most of the identified targets. Smaller proteins are easy to purify and localization information of proteins can yield insights into protein function [17,34]. A search for availability of 3D structure identified 7 proteins for which there was no structure available while for 6 of these proteins experimentally solved structures were available which points to a gap in

the structural characterization of pathogen proteins despite the fact that *C. jejuni* whole genome is sequenced. Rest of the 53 proteins had computationally solved structures available in either Modbase or ProteinModel Portal. Protein structural information can be used to a significant advantage in drug identification and validation significantly reducing the cost of high throughput experimental assays [35]. TMHMM server predicted 16 proteins that had one or more helices traversing the membrane, 6 of which were found to be antigenic above the specified threshold. ProPred-I server predicted MHC binder regions in all of the six proteins (murF, frdC, ccoP, secD, Cj1094c, tatC) for different MHC class I alleles, and these proteins represent potential vaccine candidates as most of them are transporters, surface exposed proteins. All these results are presented in Table 6.2.

6.3.4 Druggability of potential drug targets

To examine the druggability of each of the essential non-homologous protein, they were subjected to a BLASTP search against all the drug targeted proteins in DrugBank database which resulted in the identification of 34 *C. jejuni* proteins which shared high similarity to the binding partners of these drug targeted proteins from DrugBank. 9 of these 34 proteins were FDA-approved drugs or nutraceuticals while the remaining 25 were under experimentation. In Table 6.3, we have summarized the identified target protein binding partners of the all the drugs of 34 *C. jejuni* target proteins. The likelihood of being able to develop a drug-like compound to modulate the target is an important consideration that can aid drug design. The proteins for which drugs are already available can prove to be useful starting points for drug discovery.

The distribution of the essential non-homologous proteins was checked before and after similarity search against DrugBank (Figure 6.2). It was noticed that there was ~49% reduction in the number of proteins and 20% reduction in the number of pathways after DrugBank analysis; which leads to a shift in the pathway priority. Before DrugBank analysis biosynthesis of secondary metabolites, two-component system, microbial metabolism in diverse environments, peptidoglycan biosynthesis, bacterial secretion system, flagellar assembly, and lipopolysaccharide biosynthesis pathways were having maximum druggable targets but after DrugBank analysis biosynthesis of secondary metabolites, peptidoglycan biosynthesis, two-component system were the major pathways of druggable targets.

Table 6.2: Essential non-homologous proteins of *C. jejuni* with results of prioritization analysis (transmembrane and cellular location prediction)

No	Gene Name	Protein Id	Pathway Involved	Protein existence	Subcellular location	TMHMM	3D Exp	3D Model	MW (Da)
1	alr	Q9PP26	D-Alanine metabolism; Vancomycin resistance	Inferred	Cytoplasmic	No	No	Yes	37,256
2	ddl	Q9PPC2	D-Alanine metabolism; Vancomycin resistance; Peptidoglycan biosynthesis	Inferred	Cytoplasmic	No	No	Yes	39,908
3	lpxA	Q9PIM1	Lipopolysaccharide biosynthesis	Predicted	Cytoplasmic	No	Yes	Yes	28,651
4	lpxD	Q9PHU0	Lipopolysaccharide biosynthesis	Inferred	Unknown	No	No	Yes	34,679
5	lpxC	Q9PIZ5	Lipopolysaccharide biosynthesis	Inferred	Cytoplasmic	No	No	Yes	32,883
6	murA	Q9PP65	Lipopolysaccharide biosynthesis	Inferred	Cytoplasmic	No	No	Yes	45,186
7	murB	Q9PM01	Peptidoglycan biosynthesis	Inferred	Cytoplasmic	No	No	Yes	28,994
8	murC	Q9PNN7	Peptidoglycan biosynthesis	Inferred	Cytoplasmic	No	No	Yes	48,372
9	murF	Q0PA95	Peptidoglycan biosynthesis; Vancomycin resistance	Predicted	Cytoplasmic Membrane	Yes	No	Yes	55,234
10	pbpA	Q0PB07	Peptidoglycan biosynthesis; Beta-Lactam resistance	Predicted	Cytoplasmic Membrane	Yes	No	Yes	72,628
11	fba	Q0PAS0	Methane metabolism; Biosynthesis of secondary metabolites; Microbial metabolism in diverse environments	Predicted	Cytoplasmic	No	Yes	Yes	38,730
12	aroQ	Q9PJ53	Biosynthesis of secondary metabolites	Predicted	Cytoplasmic	No	No	Yes	17,595
13	pfs	Q0PC20	Biosynthesis of secondary metabolites	Predicted	Cytoplasmic	No	No	Yes	25,225
14	panD	Q9PIK3	Biosynthesis of secondary metabolites	Predicted	Cytoplasmic	No	Yes	Yes	13,974
15	panC	Q9PIK2	Biosynthesis of secondary metabolites	Predicted	Cytoplasmic	No	Yes	Yes	32,087
16	panB	Q9PIK1	Biosynthesis of secondary metabolites	Inferred	Cytoplasmic	No	No	Yes	30,202
17	pheA	Q0PBJ3	Biosynthesis of secondary metabolites	Predicted	Cytoplasmic	No	No	Yes	40,477
18	aroE	Q0PBA5	Biosynthesis of secondary metabolites	Inferred	Cytoplasmic	No	No	Yes	29,978
19	trpC	Q9PI11	Biosynthesis of secondary metabolites	Inferred	Cytoplasmic	No	No	Yes	29,573
20	ispH	P0C632	Biosynthesis of secondary metabolites	Inferred	Cytoplasmic	No	No	No	31,551
21	aroA	P0C630	Biosynthesis of secondary metabolites	Inferred	Cytoplasmic	No	No	Yes	47,288
22	aroB	Q9PNT2	Biosynthesis of secondary metabolites	Inferred	Cytoplasmic	No	No	Yes	39,596
23	dxr	Q9PMV3	Biosynthesis of secondary metabolites	Inferred	Cytoplasmic	No	No	Yes	39,548
24	Cj1370	Q0P8N9	Biosynthesis of secondary metabolites	Predicted	Cytoplasmic	No	No	Yes	17,035
25	aroC	Q9PM41	Biosynthesis of secondary metabolites	Predicted	Cytoplasmic	No	Yes	Yes	39,252
26	dapB	Q9PIT2	Biosynthesis of secondary metabolites; Microbial metabolism in diverse environments	Inferred	Cytoplasmic	No	No	Yes	26,696
27	ripB	Q0P9X4	Biosynthesis of secondary metabolites; Microbial metabolism in diverse environments	Predicted	Cytoplasmic	No	No	Yes	16,293
28	pbpB	Q0PAZ1	Peptidoglycan biosynthesis; Beta-Lactam resistance	Predicted	Cytoplasmic Membrane	Yes	No	Yes	68,378
29	pbpC	Q0PAL6	Peptidoglycan biosynthesis; Beta-Lactam resistance	Predicted	Cytoplasmic Membrane	Yes	No	Yes	67,971
30	frdC	Q0PBA2	Biosynthesis of secondary metabolites; Microbial metabolism in diverse environments; Two-component system	Predicted	Cytoplasmic Membrane	Yes	No	Yes	30,028

(continued)

Table 6.2 (CONTINUED)

No	Gene Name	Protein Id	Pathway Involved	Protein existence	Subcellular location	TMHMM	3D Exp	3D Model	MW (Da)
31	ccoP	Q0P8D0	Two-component system	Inferred	Periplasmic	Yes	No	Yes	31,162
32	cheA	Q0PBM2	Two-component system; Bacterial chemotaxis	Predicted	Cytoplasmic	No	No	Yes	85,300
33	cheV	Q0PBM1	Two-component system; Bacterial chemotaxis	Predicted	Cytoplasmic	No	No	Yes	35,858
34	cheY	P0C635	Two-component system; Bacterial chemotaxis	Inferred	Cytoplasmic	No	No	Yes	14,437
35	kdtA	Q0PAH2	Lipopolysaccharide biosynthesis	Predicted	Cytoplasmic	Yes	No	Yes	45,217
36	gmhB	Q6TG07	Lipopolysaccharide biosynthesis	Inferred	Cytoplasmic	No	No	Yes	20,246
37	waaC	Q0P9C1	Lipopolysaccharide biosynthesis	Predicted	Cytoplasmic	No	No	Yes	39,350
38	Cj0462	Q0PB51	Biosynthesis of secondary metabolites	Predicted	Cytoplasmic	No	No	Yes	39,759
39	ubiD	Q0PAX0	Biosynthesis of secondary metabolites	Predicted	Cytoplasmic	No	No	Yes	69,160
40	ispG	Q0PPM1	Biosynthesis of secondary metabolites	Inferred	Cytoplasmic	No	No	Yes	39,366
41	Cj0837c	Q0PA53	Biosynthesis of secondary metabolites	Predicted	Cytoplasmic	No	No	No	36,994
42	hisH1	Q0P8U2	Biosynthesis of secondary metabolites	Predicted	Cytoplasmic	No	No	Yes	22,699
43	Cj1368	Q0P8P1	Biosynthesis of secondary metabolites	Predicted	Cytoplasmic	No	No	Yes	41,278
44	hisB	Q0PM76	Biosynthesis of secondary metabolites	Inferred	Cytoplasmic	No	No	Yes	39,580
45	oorD	Q0PAY1	Microbial metabolism in diverse environments	Predicted	Unknown	No	No	Yes	11,443
46	oorA	Q0PAY0	Microbial metabolism in diverse environments	Predicted	Cytoplasmic	No	No	Yes	41,052
47	oorB	Q0PAX9	Microbial metabolism in diverse environments	Predicted	Cytoplasmic	No	No	Yes	31,208
48	oorC	Q0PAX8	Microbial metabolism in diverse environments	Predicted	Cytoplasmic	No	Yes	Yes	20,097
49	lysC	Q0PAT5	Biosynthesis of secondary metabolites; Microbial metabolism in diverse environments	Inferred	Unknown	No	No	Yes	42,676
50	csrA	Q0P9F1	Two-component system	Predicted	Unknown	No	No	Yes	8,441
51	Cj1153	Q0P9A3	Microbial metabolism in diverse environments;	Predicted	Periplasmic	Yes	No	Yes	10,838
52	ccoN	Q0P8C7	Two-component system	Inferred	Cytoplasmic Membrane	Yes	No	Yes	55,888
53	ccoQ	Q0P8C9	Two-component system	Predicted	Unknown	Yes	No	No	10,376
54	fliE	Q0PHY8	Flagellar assembly	Inferred	Unknown	No	No	No	10,830
55	flgA	Q0PAC1	Flagellar assembly	Predicted	Cytoplasmic	No	No	Yes	25,114
56	flaD	Q0PA11	Flagellar assembly	Predicted	Extracellular	No	No	Yes	82,051
57	flaB	P56964	Two-component system; Flagellar assembly	Inferred	Extracellular	No	No	Yes	59,186
58	flaA	P56963	Two-component system; Flagellar assembly	Inferred	Extracellular	No	No	Yes	59,039
59	flaC	P96747	Two-component system; Flagellar assembly	Predicted	Periplasmic	No	No	Yes	26,606
60	secD	Q0P9G1	Bacterial secretion system	Inferred	Cytoplasmic Membrane	Yes	No	Yes	57,251
61	secF	Q0P9G2	Bacterial secretion system	Inferred	Cytoplasmic Membrane	Yes	No	Yes	36,112
62	secE	Q0PB41	Bacterial secretion system	Inferred	Cytoplasmic Membrane	Yes	No	No	6,745
63	secY	Q0P7U3	Bacterial secretion system	Inferred	Cytoplasmic Membrane	Yes	No	Yes	46,050
64	Cj1094c	Q0P9G0	Bacterial secretion system	Predicted	Cytoplasmic Membrane	Yes	No	Yes	10,172
65	tatB	Q0PHT7	Bacterial secretion system	Inferred	Cytoplasmic	No	No	No	15,746
66	tatC	Q0PHT8	Bacterial secretion system	Inferred	Cytoplasmic Membrane	Yes	No	No	27,808

Table 6.3: Essential non-homologous proteins of *C. jejuni* inferred as similar to binding partners of drugs available in Drug Bank by homology search

No	Uniprot Id	Gene Name	Protein Name	Pathway Involved	Drug Bank Id	Drug Name	Drug Group
1	Q9PP26	alc	Alanine racemase	D-Alanine metabolism, Vancomycin resistance	DB01993,DB02142,DB03097,DB03327,DB03579,DB03766,DB03801,DB04467,DB03252,DB03801,DB00260	N-(5'-Phosphoryridoxyl)-D-Alanine, Pyridoxamine-5'-Phosphate, Prop-Hydroxyisoxazole, Pyridoxamine-5'-Phosphate-Hydroxyisoxazole, [(3-Hydroxy-Methyl)-5'-Phosphonoxy-Methyl-Pyridin-4-Ylmethyl]-Amino]-Ethyl]-Phosphonic Acid, Pyridoxyl-N,O-Cycloerylamide-5'-Monophosphate, Propanoic Acid, Lysine N α -Carboxylic Acid, Pyridoxyl-Alanine-5'-Phosphate, D-Lysine, Lysine N α -Carboxylic Acid, Cycloserine	Experimental (11); Approved
2	Q9PPC2	ddl	D-alanine-D-alanine ligase	D-Alanine metabolism, Vancomycin resistance, Peptidoglycan biosynthesis	DB00260,DB07805	Cycloserine, 3-chloro-2,2-dimethyl-n-[4-(trifluoromethyl)phenyl]propanamide	Approved; Experimental
3	Q9PIM1	lpxA	Acyl-[acyl-carrier-protein]-UDP-N-acetylglucosamine 6-O-acetyltransferase	Lipopolysaccharide biosynthesis	DB01694,DB08538	D-tartaric acid, 2-hydroxymethyl-6-octylsulfonyl-tetrahydro-pyran-3,4,5-triol	Experimental (2)
4	Q9PHU0	lpxD	UDP-3-O-acetylglucosamine N-acetyltransferase	Lipopolysaccharide biosynthesis	DB01694,DB08538	D-tartaric acid, 2-hydroxymethyl-6-octylsulfonyl-tetrahydro-pyran-3,4,5-triol	Experimental (2)
5	Q9PIZ5	lpxC	UDP-3-O-(3-hydroxymyristoyl)-N-acetylglucosamine 6-deacetylase	Lipopolysaccharide biosynthesis	DB07861,DB01991,DB04257,DB07355,DB07536,DB08231	(2R)-N-hydroxy-3-naphthalen-2-yl-2-[(naphthalen-2-ylsulfonyl)amino]propanamide, Tu-514, Palmitoleic Acid, 3-(heptyloxy)benzoic acid, N-[(1S,2R)-2-hydroxy-1-[(hydroxyamino)carbonyl]propyl]-4-[(4-(morpholin-4-ylmethyl)phenyl)ethyl]benzamide, myristic acid	Experimental (6)
6	Q9PP65	murA	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	Peptidoglycan biosynthesis	DB01879,DB02435,DB02995,DB03089,DB04174,DB04474,DB03397,DB00828,	(S)-2-[(Methyl-[2-(Naphthalene-2-Sulfonylamino)-5-(Naphthalene-2-Sulfonyloxy)-Benzoyl]-Amino)-Succinic acid, Aminomethylcyclohexane, Cyclohexylammonium Ion, L-Iso-Aspartate, 3'-I-Carboxy-1-Phosphonoxy-Ethoxy-Uridine-Diphosphate-N-Acetylglucosamine, 1-Anilino-8-Naphthalene Sulfonate, Uridine-Diphosphate-N-Acetylglucosamine, Fosfomicin	Experimental (7); Approved
7	Q9PM01	murB	UDP-N-acetylglucosamine 6-phosphate 1-uridylyltransferase	Peptidoglycan biosynthesis	DB03147	Flavin adenine dinucleotide	Approved
8	Q9PNN7	murC	UDP-N-acetylmuramate-L-alanine ligase	Peptidoglycan biosynthesis	DB01673,DB03909,DB04395	Uridine-5'-Diphosphate-N-Acetylmuramoyl-L-Alanine, Adenosine-5'-[Beta, Gamma-Methylene]Triphosphate, Phosphoraminophosphonic Acid-Adenylate Ester	Experimental (3)
9	Q9PA95	murF	UDP-N-acetylmuramoyl-tripeptide D-alanyl-D-alanine ligase	Peptidoglycan biosynthesis, Vancomycin resistance	DB06970	2-chloro-n-(3-cyano-5,6-dihydro-4h-cyclopenta[b]thiophen-2-yl)-5-diethylsulfamoyl-benzamide	Experimental

(continued)

Table 6.3 (CONTINUED)

No	Uniprot Id	Gene Name	Protein Name	Pathway Involved	Drug Bank Id	Drug Name	Drug Group
10	Q0PAS0	fla	Fructose-bisphosphate aldolase	Methane metabolism, Biosynthesis of secondary metabolites, Microbial metabolism in diverse environments	DB03026	Phosphoglycolohydroxamic Acid	Experimental
11	Q9PIK3	panD	Aspartate 1-decarboxylase	Biosynthesis of secondary metabolites	DB02175, DB03382,	Malonic acid, S-Oxy Cysteine	Experimental (2)
12	Q0PB07	pbpA	Penicillin-binding protein	Peptidoglycan biosynthesis, beta-Lactam resistance	DB00229, DB00267, DB00301, DB00417, DB00456, DB00713, DB00833, DB00948, DB01000, DB01060, DB01061, DB01066, DB01112, DB01139, DB01147, DB01602, DB01604, DB01605, DB01150, DB00415, DB00438, DB00485, DB00493, DB00567, DB00607, DB01331, DB01603, DB08795, DB00274, DB00430, DB01327, DB01328, DB01329, DB01332, DB01333, DB01414, DB01415, DB01598, DB04570, DB04133, DB03313, DB08375, DB00447, DB01163, DB05659, DB00739, DB01140, DB00303	Cefotiam, Cefmenoxime, Flucloxacillin, Penicillin V, Cefalotin, Oxacillin, Cefaclor, Mezlocillin, Cycloacillin, Amoxicillin, Azlocillin, Cefitoren, Cefuroxime, Cefapirin, Cloxacillin, Bacampicillin, Pivampicillin, Pivmecillinam, Cefprozil, Ampicillin, Cefazidime, Dicloxacillin, Cefotaxime, Cephalixin, Nafcillin, Cefoxitin, Meticillin, Azidocillin, Cefmetazole, Cefpiramide, Cefazolin, Cefonicid, Cefoperazone, Cefizoxime, Cefradine, Cefacetrile, Cefibuten, Imipenem, Latamoxef, Degraded Cephaloridine, Cephalosporin C, Penotaxime group, Loracarbef, Amdinocillin, faropenem medoxomil, Hetacillin, Cefadroxil, Ertapenem	Approved (39); Experimental (3); Withdrawn (2); Investigational; Approved, withdrawn (2); Approved, investigational
13	Q9PI53	aroQ	3-dehydroquinate dehydratase	Biosynthesis of secondary metabolites	DB03868, DB04698, DB02801, DB03739, DB02786, DB04347, DB04656, DB08485	3-Dehydroquinate Acid, N-(1,4-dihydro-5h-tetrazol-5-ylidene)-9-oxo-9h-xanthene-2-sulfonamide, 2,3-Anhydro-Quinic Acid, 3-Hydroxyimino Quinic Acid, 2-Anhydro-3-Fluoro-Quinic Acid, 3-Dehydroshikimate, 1,3,4-trihydroxy-5-(3-phenoxypyrrolyl)-cyclohexane-1-carboxylic acid, (1S,4S,5S)-1,4,5-trihydroxy-3-[3-(phenylthio)phenyl]cyclohex-2-ene-1-carboxylic acid	Experimental (8)
14	Q0PC20	pfs	Aminodeoxyfutal osine nucleosidase	Biosynthesis of secondary metabolites	DB07463, DB02158, DB02281, DB02933, DB08606, DB07649, DB02857, DB00173	(3R,4S)-1-[(4-amino-5H-pyrrolo[3,2-d]pyrimidin-7-yl)methyl]-4-[(butylsulfonyl)methyl]pyrrolidin-3-ol, (1S)-1-(9-Deazaadenin-9-Yl)-1,4,5-Triideoxy-1,4-Amino-5-Methylthio-D-Ribitol, Formycin, 5'-Deoxy-5'-(Methylthio)-Tubercidin, (3R,4S)-1-[(4-amino-5H-pyrrolo[3,2-d]pyrimidin-7-yl)methyl]-4-[(methylsulfonyl)methyl]pyrrolidin-3-ol, (3R,4S)-1-[(4-amino-5H-pyrrolo[3,2-d]pyrimidin-7-yl)methyl]-4-[(benzylsulfonyl)methyl]pyrrolidin-3-ol, Guanosine, Adenine	Experimental (7); Approved, Nutritional
15	Q9PIK2	panC	Pantothenate synthetase	Biosynthesis of secondary metabolites	DB01930, DB02596, DB02694, DB03107	2,4-Dihydroxy-3,3-Dimethyl-Butyrate, Alpha.Beta-Methyleadenosine-5'-Triphosphate, Pantoyl Adenylate, Betn-Alanine	Experimental (4)
16	Q9PIK1	panB	3-methyl-2-oxobutanoate hydroxymethyltransferase	Biosynthesis of secondary metabolites	DB03795, DB04074	2-Dehydropanthoate, Alpha-ketoisovalerate	Experimental (2)

(continued)

Table 6.3 (CONTINUED)

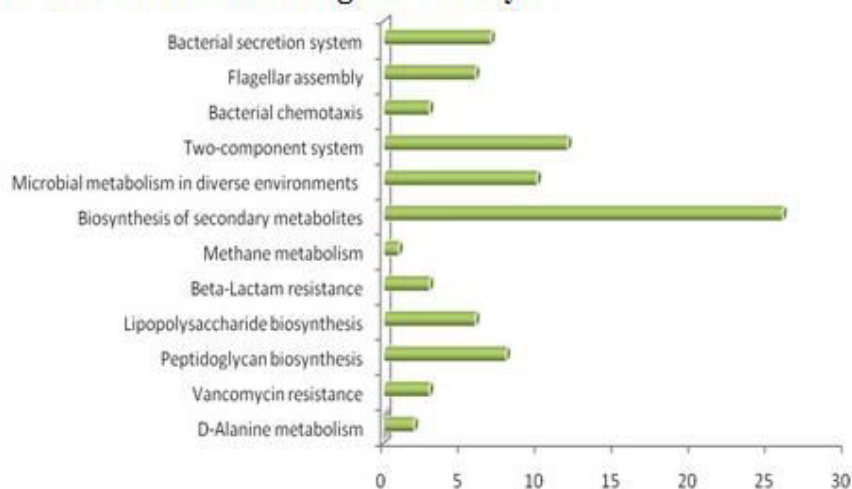
No	Uniprot Id	Gene Name	Protein Name	Pathway Involved	Drug Bank Id	Drug Name	Drug Group
17	Q0PBU3	pheA	Chorismate mutase/prephenate dehydratase	Biosynthesis of secondary metabolites	DB08648, DB03884	8-Hydroxy-2-oxa-bicyclo[3.3.1]non-6-ene-3,5-dicarboxylic acid, 3-Phenylpyruvic Acid	Experimental (2)
18	Q0PBA5	aroE	Shikimate dehydrogenase	Biosynthesis of secondary metabolites	DB02363, DB03461, DB04447	2'-Monophosphoadenosine-5'-Diphosphate, 2'-Monophosphoadenosine 5'-Diphosphoribose, 1,4-Dithiothreitol	Experimental (3)
19	Q9P111	trpC	Indole-3-glycerol phosphate synthase	Biosynthesis of secondary metabolites	DB03543	1-(O-Carboxy-Phenylamino)-1-Deoxy-D-Ribulose-5-Phosphate	Experimental
20	P0C632	ispH	4-Hydroxy-3-methylbut-2-enyl diphosphate reductase	Biosynthesis of secondary metabolites	DB01785, DB04714	Dimethylallyl Diphosphate, Isopentenyl Pyrophosphate	Experimental (2)
21	P0C630	aroA	3-Phosphoshikimate 1-carboxyvinyltransferase	Biosynthesis of secondary metabolites	DB04328, DB04539, DB01942, DB03116	Shikimate-3-Phosphate, Glyphosate, Formic Acid, 5-(1-Carboxy-1-Phosphonoxy-Ethoxy)-Shikimate-3-Phosphate	Experimental (4)
22	Q9PNT2	aroB	3-dehydroquinate synthase	Biosynthesis of secondary metabolites	DB02592	Carbaphosphonate	Experimental
23	Q9PMV3	dxr	1-deoxy-D-xylulose 5-phosphate reductoisomerase	Biosynthesis of secondary metabolites	DB02496, DB02948, DB03649, DB04272	1-Deoxy-D-xylulose 5-phosphate, Fosmidomycin, [(5-Chloro-2-Pyridinyl)Amino] Methylene]-1,1-Bisphosphonate, Citric Acid	Experimental (3); Nutraaceutical
24	Q0P8N9	Cj1370		Biosynthesis of secondary metabolites	DB02075	(1s)-1-(9-Deazahypoxanthin-9-yl)-4,4-Dideoxy-1,4-Imino-D-Ribitol-5-Phosphate	Experimental
25	Q9PM41	aroC	Chorismate synthase	Biosynthesis of secondary metabolites	DB03247, DB03350, DB03969, DB04267	Riboflavin Monophosphate, Cobalt Hexamine Ion, 3-Acetyl Pyridine Adenine Dinucleotide, Dipicolinic Acid	Experimental (4)
26	Q9PTT2	dapB	4-Hydroxy-tetrahydrodipicolinate reductase	Biosynthesis of secondary metabolites, Microbial metabolism in diverse environments	DB03969, DB04267	3-Acetyl Pyridine Adenine Dinucleotide, Dipicolinic Acid	Experimental (2)
27	Q0P9X4	rpiB	Ribose 5-phosphate isomerase	Biosynthesis of secondary metabolites, Microbial metabolism in diverse environments	DB03661, DB03108, DB04496	Cysteinesulfonic Acid, 4-Phospho-D-Erythronate, 4-Phospho-D-Erythronolhydroxamic Acid	Experimental (3)

(continued)

Table 6.3 (CONTINUED)

No	UniProt Id	Gene Name	Protein Name	Pathway Involved	Drug Bank Id	Drug Name	Drug Group
28	Q0PAZ1	pbpB	Pelative penicillin-binding protein	Peptide:glycan biosynthesis, beta-Lactam resistance	DB00274, DB00430, DB00438, DB01327, DB01328, DB01329, DB01331, DB01332, DB01414, DB01415, DB01598, DB04570, DB00319, DB00415, DB00456, DB00485, DB00493, DB00567, DB00607, DB00713, DB01603, DB08795, DB00303, DB08401, DB08313, DB08508, DB00739, DB01140, DB01163	Cefmetazole, Cefpiramide, Cefazidime, Cefazolin, Cefonicid, Cefoperazone, Cefoxitin, Cefizoxime, Cefaclor, Cefibuten, Imipenem, Latamoxef, Piperacillin, Ampicillin, Cefalotin, Dicloxacillin, Ceftriaxime, Cephalexin, Nafcillin, Oxacillin, Meticillin, Azticillin, Ertapenem, (2e)-2-((2s)-2-carboxy-2-[[phenoxyacetyl(aminio)ethoxy]iminopentenedioic acid, Cephalosporin C, N-benzoyl-D-alanine, Hetsacillin, Cefadroxil, Amcinocillin	Approved (22); Approved, Investigational; Experimental (3); Approved, withdrawn (2); Withdrawn,
29	Q0PAL6	pbpC	Penicillin-binding protein	Peptide:glycan biosynthesis, beta-Lactam resistance	DB01327	Cefazolin	Approved
30	Q0PBA2	frdC	Fumarate reductase cytochrome B subunit	Biosynthesis of secondary metabolites, Microbial metabolism in diverse environments, Two-component system	DB07569	2,3-dimethyl-1,4-naphthoquinone	Experimental
31	Q0PSD0	ecpP	Cbb3-type cytochrome c oxidase subunit	Two-component system	DB02949, DB03224	2-Acetyl-Protoporphyrin IX, 2-Formyl-Protoporphyrin IX	Experimental (2)
32	Q0PBM2	cleA	Chemotaxis hisidine kinase	Two-component system, Bacterial chemotaxis	DB02524, DB03909, DB04395	Spino(2,4,6-Tinitrobenzene)[2,2a]2,3,4-Methylene-Adenine-Triphosphate, Adenosine-5'-[Beta, Gamma-Methylene]Triphosphate, Phosphoaminophosphonic Acid-Adenyate Ester	Experimental (3)
33	Q0PBM1	cleV	Chemotaxis protein	Two-component system, Bacterial chemotaxis	DB02355, DB02596, DB07706, DB02461, DB02487, DB04156	Adenosine-5'-R-p-Alpha-Thio-Triphosphate, Alpha.Beta-Methyleneadenosine-5'-Triphosphate, 2,3,17beta-trihydroxy-1,3,5(10)-estratriene, S-Methyl Phosphocysteine, 3-Aminosuccinimide, Aspartate Beryllium Trifluoride	Experimental (6)
34	P0C635	cleV	Chemotaxis protein	Two-component system, Bacterial chemotaxis	DB02461, DB03487, DB04156,	S-Methyl Phosphocysteine, 3-Aminosuccinimide, Aspartate Beryllium Trifluoride,	Experimental (3)

Protein Hits Before Drug Bank Analysis



Protein Hits After Drug Bank Analysis

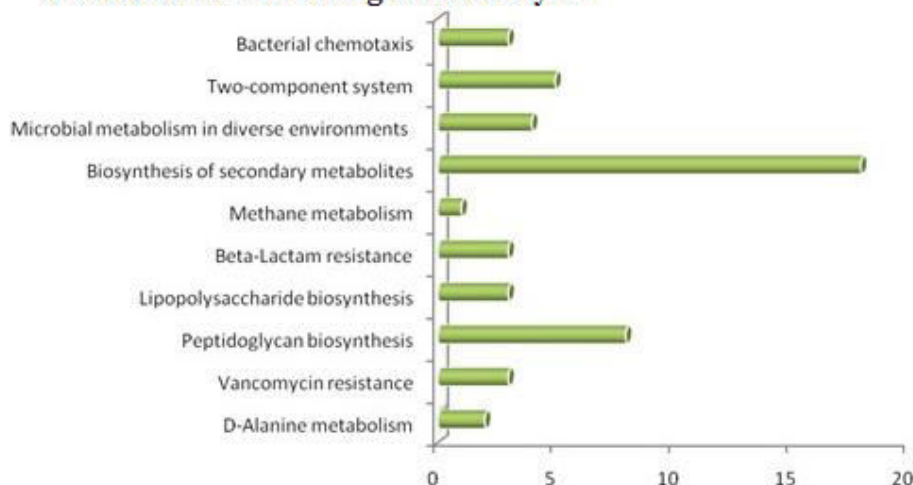


Figure 6.2: Distribution of essential non-homologous proteins of *C. jejuni* in different pathways before and after drug bank search for druggability potential evaluation.

Finally, we have narrowed down the search of therapeutic targets to final 14 prioritized drug targets and vaccine candidates. 5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase (pfs) enzyme catalyzes the direct conversion of aminodeoxyfutalosine (AFL) into dehypoxanthine futalosine (DHFL) and adenine via the hydrolysis of the N-glycosidic bond which represents an essential step in menaquinone biosynthesis pathway [36]. This enzyme is an attractive drug target as it is involved in many pathways such as ubiquinone and other terpenoid-quinone biosynthesis, cysteine and methionine metabolism, biosynthesis of amino acids.

Both alanine racemase (alr) and pyridoxal 5'-phosphate-dependent enzyme that catalyze the interconversion of L-alanine and D-alanine and D-alanyl-alanine synthetase A (ddl) involved in cell wall formation by joining two of the D-alanine residues together catalyzing the formation of the ATP-dependent D-alanine-D-alanine dipeptide bond between the resulting D-alanine molecules. Inhibition of these two enzymes leads to effective inhibition of peptidoglycan synthesis and ddl has been proposed to be an attractive drug target in *Mycobacterium tuberculosis* [37]. Vancomycin binds to ddl, a peptidoglycan precursor under normal conditions forming a stable complex and inhibits cell wall synthesis [38] ultimately leading to cell lysis, thus playing an important role in vancomycin resistance pathway. Penicillin Binding Proteins (PBPs) are of special interest as these are target site for beta-lactam antibiotics. They also play an important role in cell wall formation. pbpA appear to be important for cell division and essential for growth [39]. pbpB also is critical for bacterial growth and cell wall biosynthesis [40]. pbpC is a major protein of cell division complex. PBPs have already been utilized as model drug target system [41]. PBPs were found to be highly similar to the binding partners of many FDA-approved and experimental drugs. Hence PBPs can be considered of high potential for experimental validation as vaccine candidates.

UDP-N-acetylglucosamine 1-carboxyvinyltransferase (murA) and UDP-N-acetylenolpyruvoylglucosamine reductase (murB), both catalyze important reactions in the peptidoglycan precursor synthesis. murA catalyzes transfer of an enolpyruvate residue from phosphoenolpyruvate (PEP) to position 3 of UDP-N-acetylglucosamine followed by a MurB-catalysed NADPH dependent reduction of the UDP-N-acetylglucosamine enolpyruvate to UDP-N-acetylmuramic acid. Majority of antibiotics in clinical use today target later steps of peptidoglycan synthesis [42]. murB has previously been reviewed to be essential in *Escherichia coli* [43]. murA to murF genes are all essential and highly conserved among bacterial species thus holding a great promise as future therapeutic drug targets. Both frdC (Fumarate reductase cytochrome B subunit) and ccoP are important constituents of oxidative phosphorylation pathway for ATP formation, often called as molecular unit of energy transfer. While frdC couples the reduction of fumarate to succinate with the oxidation of quinol to quinone, ccoP (Cbb3-type cytochrome c oxidase subunit) is required for transfer of electrons from donor cytochrome c via its heme groups to CcoO subunit. secD (Protein translocase subunit SecD) a part of the Sec protein translocase complex, tatC (Sec-independent protein translocase protein

TatC) an important part of the twin-arginine translocation (Tat) system and Cj1094c (Putative preprotein translocase protein) transports large proteins across membranes. Twin-arginine translocation (TAT) pathway is important to bacterial growth and virulence [44,45]. secD and Cj1094c help in secretion across the inner membrane via preprotein translocase pathway. Transport proteins are associated with pathogenesis and virulence and have been identified as potential vaccine candidates in several previous studies as well [46,47]. Peptidoglycan is an important component of bacterial cell wall responsible for maintaining a definite cell shape and primarily conferring mechanical resistance to higher osmotic pressure [48]. Any interference with peptidoglycan biosynthesis will result in cell lysis. Peptidoglycan biosynthesis, the pathway with largest distribution of final identified drug targets can be exploited for therapeutic drug targets owing to its multiple target enzymes whose inhibition could lead to disruption of cell wall and in turn attenuate bacterial cell growth.

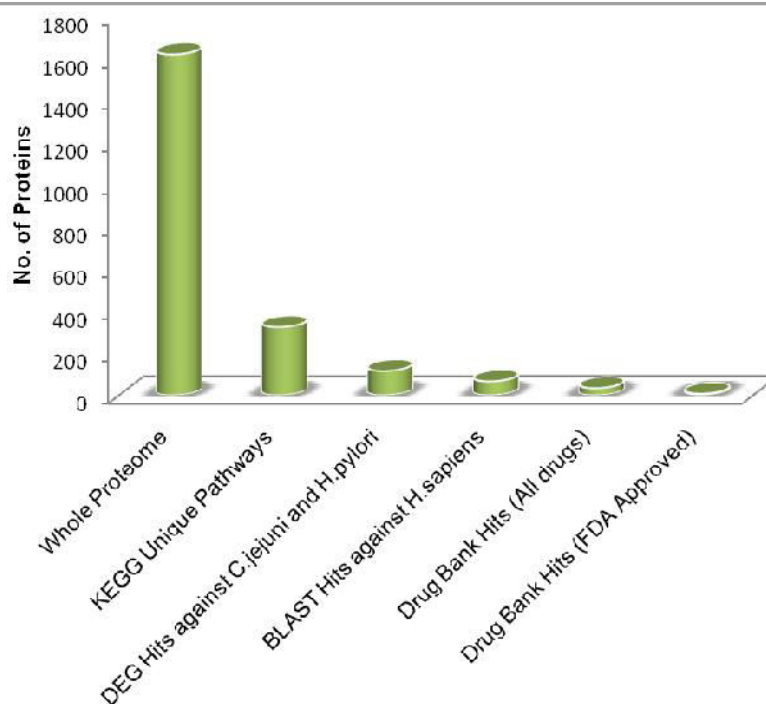


Figure 6.3: Step-wise reduction in the total no of proteins in the subtractive analysis of *C. jejuni* proteome for drug target identification.

The computational subtractive genomics approach integrated with comparative pathway analysis resulted in a significant reduction in the number of protein targets (Figure 6.3) at each step. Thus

we were able to identify several essential proteins critical for bacterial growth and survival and with minimum toxicity to host, that can be targeted for effective drug design to combat *Campylobacter* infections.

Conclusion

We have performed subtractive genomics analyses of the *C. jejuni* pathogenic strain NCTC11168, and have identified several proteins in the genome that can prove to be potential targets for effective drug design. As many of the identified drug targets have already been reviewed to play critical role in the metabolic pathways that regulate bacterial growth and survival, a systematic approach to develop antibiotics against the identified targets would likely be very promising for the treatment of *Campylobacter* infections. The information about these targets can also lead to significant progress in testing the efficacy of already existing drugs which is as equally important as development of new drugs. It is believed that the drugs developed against these identified targets will be pathogen specific and with minimal toxic effects on host.

References

1. WHO. *Campylobacter* Fact Sheet.(World Health Organization, 2011)
2. CDC. Preliminary FoodNet data on the incidence of foodborne illnesses--selected sites, United States, 1999. *MMWR. Morbidity and mortality weekly report*, 49(10), 201-205 (2000).
3. CDC. Antibiotic Resistance Threats in the United States, 2013. (Centers for Disease Control and Prevention, 2013)
4. Akhtar SQ. Antimicrobial sensitivity and plasmid-mediated tetracycline resistance in *Campylobacter jejuni* isolated in Bangladesh. *Chemotherapy*, 34(4), 326-331 (1988).
5. Engberg J, Aarestrup FM, Taylor DE, Gerner-Smidt P, Nachamkin I. Quinolone and macrolide resistance in *Campylobacter jejuni* and *C. coli*: resistance mechanisms and trends in human isolates. *Emerging Infectious Diseases*, 7(1), 24-34 (2001).
6. Hoge CW, Gambel JM, Srijan A, Pitarangsi C, Echeverria P. Trends in Antibiotic Resistance Among Diarrheal Pathogens Isolated in Thailand Over 15 Years. *Clinical Infectious Diseases*, 26(2), 341-345 (1998).
7. Reina J, Ros MJ, Serra A. Susceptibilities to 10 antimicrobial agents of 1,220 *Campylobacter* strains isolated from 1987 to 1993 from feces of pediatric patients. *Antimicrobial agents and chemotherapy*, 38(12), 2917-2920 (1994).
8. Fleischmann RD, Adams MD, White O *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science (New York, N.Y.)*, 269(5223), 496-512 (1995).

9. Fraser CM, Gocayne JD, White O *et al.* The minimal gene complement of *Mycoplasma genitalium*. *Science (New York, N.Y.)*, 270(5235), 397-403 (1995).
10. Chawley P, Samal HB, Prava J, Suar M, Mahapatra RK. Comparative genomics study for identification of drug and vaccine targets in *Vibrio cholerae*: MurA ligase as a case study. *Genomics*, 103(1), 83-93 (2013).
11. Ghosh S, Prava J, Samal HB, Suar M, Mahapatra RK. Comparative genomics study for the identification of drug and vaccine targets in *Staphylococcus aureus*: MurA ligase enzyme as a proposed candidate. *Journal of microbiological methods*, 101, 1-8 (2014).
12. Samal HB, Prava J, Suar M, Mahapatra RK. Comparative genomics study of *Salmonella* Typhimurium LT2 for the identification of putative therapeutic candidates. *Journal of theoretical biology*, 369C, 67-79 (2015).
13. Sarangi AN, Aggarwal R, Rahman Q, Trivedi N. Subtractive genomics approach for in silico identification and characterization of novel drug targets in *Neisseria meningitidis* serogroup B. *Journal of Computer Science & Systems Biology*, 2(5), 255-258 (2009).
14. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), 27-30 (2000).
15. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403-410 (1990).
16. Zhang R, Ou H-Y, Zhang C-T. DEG: a database of essential genes. *Nucleic acids research*, 32(Database issue), D271-272 (2004).
17. Yu NY, Wagner JR, Laird MR *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics (Oxford, England)*, 26(13), 1608-1615 (2010).
18. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*, 305(3), 567-580 (2001).
19. Berman HM, Westbrook J, Feng Z *et al.* The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235-242 (2000).
20. Pieper U, Eswar N, Davis FP *et al.* MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic acids research*, 34(Database issue), D291-295 (2006).
21. Haas J, Roth S, Arnold K *et al.* The Protein Model Portal--a comprehensive resource for protein structure and model information. *Database : the journal of biological databases and curation*, 2013(0), bat031-bat031 (2013).
22. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC bioinformatics*, 8(1), 4-4 (2007).
23. Singh H, Raghava GPS. ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics (Oxford, England)*, 19(8), 1009-1014 (2003).

24. Wishart DS, Knox C, Guo AC *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(Database issue), D901-906 (2008).
25. Aguerro F, Al-Lazikani B, Aslett M *et al.* Genomic-scale prioritization of drug targets: the TDR Targets database. *Nature reviews. Drug discovery*, 7(11), 900-907 (2008).
26. Winzeler EA, Shoemaker DD, Astromoff A *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science (New York, N.Y.)*, 285(5429), 901-906 (1999).
27. Titus RG, Gueiros-Filho FJ, de Freitas LA, Beverley SM. Development of a safe live *Leishmania* vaccine line by gene replacement. *Proceedings of the National Academy of Sciences of the United States of America*, 92(22), 10267-10271 (1995).
28. Butt AM, Nasrullah I, Tahir S, Tong Y. Comparative genomics analysis of *Mycobacterium ulcerans* for the identification of putative essential genes and therapeutic candidates. *PloS one*, 7(8), e43080-e43080 (2012).
29. Doyle MA, Gasser RB, Woodcroft BJ, Hall RS, Ralph SA. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC genomics*, 11, 222-222 (2010).
30. Volker C, Brown JR. Bioinformatics and the discovery of novel anti-microbial targets. *Current drug targets. Infectious disorders*, 2(4), 279-290 (2002).
31. Tilton F, Priya L, Nair K, Nair A. *insilico* metabolic pathway analysis for potential drug target in *Campylobacter jejuni*. *International Journal of Healthcare and Pharmaceutical Research*, 3(3), 029-032 (2014).
32. Perriere G, Thioulouse J. Use and misuse of correspondence analysis in codon usage studies. *Nucleic acids research*, 30(20), 4548-4555 (2002).
33. Rocha EPC, Danchin A. Gene essentiality determines chromosome organisation in bacteria. *Nucleic acids research*, 31(22), 6570-6577 (2003).
34. Duffield M, Cooper I, McAlister E, Bayliss M, Ford D, Oyston P. Predicting conserved essential genes in bacteria: in silico identification of putative drug targets. *Molecular bioSystems*, 6(12), 2482-2489 (2010).
35. Grant MA. Protein structure prediction in structure-based ligand design and virtual screening. *Combinatorial chemistry & high throughput screening*, 12(10), 940-960 (2009).
36. Li X, Apel D, Gaynor EC, Tanner ME. 5'-methylthioadenosine nucleosidase is implicated in playing a key role in a modified futasine pathway for menaquinone biosynthesis in *Campylobacter jejuni*. *The Journal of biological chemistry*, 286(22), 19392-19398 (2011).
37. Prosser GA, de Carvalho LPS. Kinetic mechanism and inhibition of *Mycobacterium tuberculosis* D-alanine:D-alanine ligase by the antibiotic D-cycloserine. *The FEBS journal*, 280(4), 1150-1166 (2013).
38. Howden BP, Davies JK, Johnson PDR, Stinear TP, Grayson ML. Reduced vancomycin susceptibility in *Staphylococcus aureus*, including vancomycin-intermediate and heterogeneous vancomycin-intermediate strains: resistance mechanisms, laboratory detection, and clinical implications. *Clinical microbiology reviews*, 23(1), 99-139 (2010).
39. Wada A, Watanabe H. Penicillin-binding protein 1 of *Staphylococcus aureus* is essential for growth. *Journal of bacteriology*, 180(10), 2759-2765 (1998).

40. Pinho MG, de Lencastre H, Tomasz A. An acquired and a native penicillin-binding protein cooperate in building the cell wall of drug-resistant staphylococci. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19), 10886-10891 (2001).
41. von Rechenberg M, Blake BK, Ho Y-SJ *et al.* Ampicillin/penicillin-binding protein interactions as a model drug-target system to optimize affinity pull-down and mass spectrometric strategies for target and pathway identification. *Proteomics*, 5(7), 1764-1773 (2005).
42. El Zoeiby A, Sanschagrin Fo, Levesque RC. Structure and function of the Mur enzymes: development of novel inhibitors. *Molecular microbiology*, 47(1), 1-12 (2003).
43. Pucci MJ, Discotto LF, Dougherty TJ. Cloning and identification of the Escherichia coli murB DNA sequence, which encodes UDP-N-acetylenolpyruvoylglucosamine reductase. *Journal of bacteriology*, 174(5), 1690-1693 (1992).
44. Ding Z, Christie PJ. Agrobacterium tumefaciens twin-arginine-dependent translocation is important for virulence, flagellation, and chemotaxis but not type IV secretion. *Journal of bacteriology*, 185(3), 760-771 (2003).
45. Lavander M, Ericsson SK, Broms JE, Forsberg A. Twin arginine translocation in Yersinia. *Advances in experimental medicine and biology*, 603, 258-267 (2007).
46. Garmory HS, Titball RW. ATP-binding cassette transporters are targets for the development of antibacterial vaccines and therapies. *Infection and immunity*, 72(12), 6757-6763 (2004).
47. Harris JA, Roy K, Woo-Rasberry V *et al.* Directed evaluation of enterotoxigenic Escherichia coli autotransporter proteins as putative vaccine candidates. *PLoS neglected tropical diseases*, 5(12), e1428-e1428 (2011).
48. Vollmer W, Blanot D, De Pedro MA. Peptidoglycan structure and architecture. *FEMS Microbiology Reviews*, 32(2), 149-167 (2008).

CHAPTER 7

Tapping into *Salmonella typhimurium* LT2 genome in a quest to explore its therapeutic arsenal: a metabolic network modeling approach

7.1 Introduction

Salmonella enterica sv. *typhimurium* (*S. typhimurium*), gram negative, rod shaped bacteria are one of the most common and widely distributed bacterial causes of food-borne illness. *Salmonella*'s annual incidence rate is estimated to be tens of millions of human cases worldwide [1]. *S. typhimurium* is the classical broad-host-range serovar, since it exhibits promiscuous behavior in their ability to infect broadly diverse range of phylogenetically unrelated host species including humans, livestock, domestic fowl, rodents, and birds [2]. *Salmonella* infection is usually associated with self-limited gastroenteritis; however, it can manifest itself in highly invasive form which has surfaced as a major public health concern in sub-Saharan Africa [3,4]. *S. typhimurium* has exhibited a gradual increase in antibiotic resistance against conventional antibiotics such as ampicillin, chloramphenicol, and sulfamethoxazole-trimethoprim, fluoroquinolones and extended-spectrum cephalosporins; and emergence of multidrug resistant strains isolated from humans and animals worldwide [5-7]. Growing antibiotic resistance in nontyphoid *Salmonella* serotypes is a global concern owing to the broad-range host specificity of the pathogen [2] and its potential of conjugal transmission of antibiotic resistance to other pathogens [8]. Continuous surveillance of antimicrobial resistance and its impact on clinical medicine is imperative to *S. typhimurium* transmission monitoring. To keep pace with the continuously evolving bacteria, there is growing need to restock the antibiotic pipeline. With bioterrorism threats, growing number of drug resistant bacteria, and emergence of new infectious diseases it becomes imperative to develop new antibiotics particularly directed against multidrug resistant gram-negative bacteria in hospitals and community-acquired pathogens [9], otherwise there is a serious risk that a major proportion of infectious diseases will effectively be untreatable.

Genome-scale metabolic network models seek to decipher the complex molecular mechanisms governing physiological and biochemical aspects of an organism. These networks provide vital, cost effective framework for understanding and predicting the cellular behavior in physiological as well as pathophysiological state [10]. Genome scale reconstruction of metabolic networks allows developing system-oriented drug-design strategies by integrating biological and chemical information on genes, metabolites, biochemical reactions, drugs, diseases and drug targets [11]. Constraint-based methods can then be applied to such models to describe the pathogen metabolic

physiology quantitatively since microbial cells functioning is also limited due to governing constraints [12]. Genome-scale metabolic network modeling has been utilized for drug target identification in some pathogens [13-16].

In this work, we interrogate the metabolic network model of *S. typhimurium* to identify effective and novel drug targets by applying a system-level approach that efficiently utilizes the constraint based methods to metabolic network modeling. The drug targets were predicted employing the concept of metabolite essentiality i.e. the metabolite is critical to the bacterial growth (biomass production) and were further screened out applying additional filtration criteria. It is believed that this strategy can be effectively applied for drug targeting in other pathogens as well in a cost effective manner.

7.2 Methods

To predict potential drug candidates, we used a genome scale metabolic model of *S. Typhimurium* (iRR1083) [17]. Constraint based modeling of metabolic networks holds immense predictive powers for identification of critical genes, metabolites and reactions involved in metabolism. Constraint based modeling offers an added advantage which is the need to determine few parameters from experimental data as compared to other modeling approaches. We have used constraint based approach FBA (Flux Balance Analysis) and its extension MOMA (Minimization of Metabolic Adjustment) which determine flux distribution through the metabolic network by maximizing the objective function.

FBA: FBA formulates a metabolic network as the system of stoichiometry balanced set of equations which describe changes in the concentration as the dot product of a matrix of the stoichiometric coefficients (the stoichiometric matrix S of size $m \times n$ with m metabolites and n reactions) and the vector v of the unknown fluxes. The right-hand side of the dot product is set to zero implying steady state.

$$S \cdot v = 0$$

Since, in a metabolic network number of reactions far outcompetes the metabolites number ($n > m$), the system is under-determined (with $n-m$ degrees of freedom), entailing the imposition

of additional constraints. Such under-determined systems are then solved using Linear Programming (LP).

$$\begin{aligned} & \max f^T v \\ & \text{subject to } Sv = 0 \\ & \text{and } v_{\min} \leq v \leq v_{\max} \end{aligned}$$

The expression to be maximized ($f^T v$) is defined as objective function and the inequalities define the minimal and the maximal rates of flux for every reaction.

MOMA: MOMA accurately describes the transient growth rates immediate after gene knockouts while FBA specifically focus on final optimal metabolic steady state. MOMA utilizes Quadratic Programming (QP) to search for a point in the solution space of the mutant strain which is contiguous to the optimal point in the solution space of wild type strain. MOMA tries to minimize the sum of the squares difference between wild type and mutant flux distribution. MOMA tests the hypothesis of minimizing flux redistribution in the mutant strain and thus effectively capturing the mutant phenotype characteristics.

7.2.1 Gene knockout analysis

We carried out gene knockout studies by performing simulation of the mutant strains (model with gene deletions) for all genes of *S. typhimurium* (iRR1083) and analysis was carried out using OptFlux [18]. If the gene knockout reduces the objective function (biomass) by >90%, the corresponding gene is presumed to be indispensable. Gene deletion analysis could lend a hand to identify genes or metabolites which are indispensable for bacterial growth. An *in silico* knockout experiment by using MOMA approach was performed for each gene in the metabolic model. The results were compared against the predictions by FBA approach. Genes common to both predictions were considered as essential to the growth and maintenance of the bacterial metabolic system.

7.2.2 Identification of essential non-homologous pathogen proteins

Genes essential to bacterial growth which were obtained from *in silico* knockout studies were compared against experimentally identified essential genes of *Salmonella* spp. deposited in the

Database of Essential genes (DEG version 10.9) [19]. The genes were searched against DEG which hosts records of essential genomic elements critical for an organism's survival by using an e-value of 10^{-5} and bit score 100. The background organism against which search was performed include *Salmonella enterica* serovar *Typhi*, *Salmonella enterica* serovar *Typhi* Ty2, *Salmonella enterica* serovar *Typhimurium* SL1344, *Salmonella enterica* subsp. *Enterica* serovar *Typhimurium* str. 14028S, and *Salmonella typhimurium* LT2. The filtered gene list was further screened out on the basis of homology with the human proteome at an e-value cut off of 0.05. Genes which do not show any hits below e-value inclusion threshold were selected as therapeutic drug targets.

7.2.3 Prioritization of drug targets

The molecular and structural properties of proteins encoded by selected genes were calculated which aid in prioritization of those drug targets that will most likely lead to effective treatments. Subcellular localization of identified proteins was predicted using PSORTb [20] which uses a Bayesian network model to calculate associated probability value for five major localization sites viz. cytoplasmic, inner membrane, periplasmic, outer membrane and extracellular with p-value criteria 7.5. TMHMM server which is a Hidden Markov Model (HMM) based tool for prediction of alpha helices in membrane proteins was used for transmembrane predictions [21]. A search was performed to identify the proteins for which experimentally or computationally solved structures are available in PDB [22] or Modbase [23] or Protein Model Portal [24]. For uncharacterized proteins homology models were predicted through RaptorX server [25]; which is based on novel nonlinear context-specific alignment and probabilistic consistency algorithm to search for remote templates. A search of identified proteins in Uniprot was performed for retrieving information on their molecular weight.

7.2.4 Druggability and assayability

To evaluate the druggability potential of each identified therapeutic target, we subjected each protein to a BLASTP search against DrugBank with an e-value 0.0001. Druggability potential evaluates the ability of a particular biological target to bind with high affinity to known drugs; DrugBank is a unique comprehensive resource that integrates drug data with drug target information at sequence, structure and pathway level [26]. DrugBank version 4.2 currently has

information about 7737 drug entries including 1585 FDA-approved small molecule drugs, 158 FDA-approved biotech (protein/peptide) drugs, 89 nutraceuticals and over 6000 experimental drugs; along with 4281 non-redundant protein sequences (drug target/enzyme/transporter/carrier) linked to these drugs. Assayability includes information on the availability of biochemical and/or cellular assays and reagents. Enzyme assayability is usually hindered by difficulties in producing soluble recombinant protein.

7.2.5 Protein-Protein Interaction (PPI) network analysis

The proteins finally picked as potential drug targets were studied for their PPIs to identify the pathways enriched in these proteins and their interaction in *Salmonella* proteome. PPI networks are frequently employed in prioritization of drug targets. We used STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database [27] for analyzing PPI networks of the selected set of drug targets. STRING is a comprehensive resource of 5 million proteins and >200 million PPIs from more than a thousand organisms encompassing experimental, predicted and transferred interactions along with accessory information such as protein domains and structures. These networks present a holistic view of the interacting proteins within the pathogen proteome.

7.3 Results and discussion

The 4,857-kb chromosome (Figure 7.1) and 94-kb virulence plasmid of *Salmonella enterica* serovar *typhimurium* LT2 strain first isolated in 1940s in Sweden were sequenced in 2001. The complete genome sequence of *S. typhimurium* LT2 genome codes for 4,489 coding sequences (CDS/ORFs) including 39 pseudogenes and has GC content of 53 percent [28]. *S. typhimurium* LT2 genome is mosaic composed mostly of collinear regions interspersed with loops or islands. These islands called as *Salmonella* Pathogenicity Islands (SPI's) often code for pathogenicity traits [29,30]. The host generalist nature of *S. typhimurium* LT2 strain could be a consequence of fewer pseudogenes (39) as compared to its close relative *S. typhi* (204) [31] which is host specialist in that it is restricted to grow in humans only. Understanding pathogen metabolism has gained momentum to seek insights into host defense mechanisms and to better understand the pathogen intracellular lifestyle. The genome scale metabolic network model of *S. typhimurium* LT2 (iRR1083) containing 1,083 genes, 973 proteins and 744 metabolites which catalyze 1087 reactions was used for identification of drug targets in *Salmonella*.

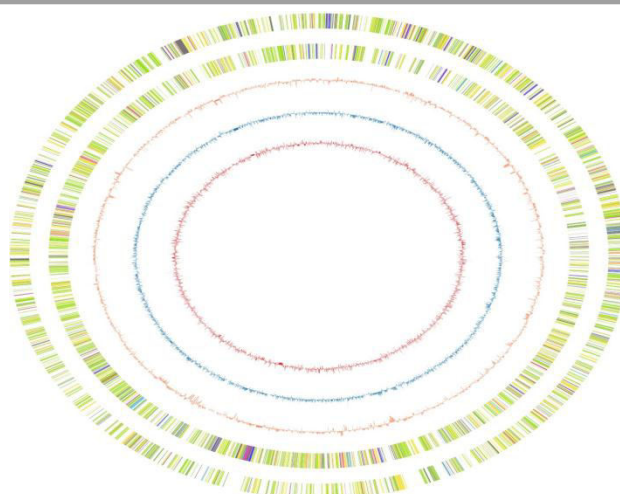


Figure 7.1: Circular genome of *S. typhimurium* LT2

(i) ORFs plus strand: The outer most ring shows ORFs plus strand. (ii) The second ring indicates the ORFs negative strand. In two outermost rings green, yellow, purple, light blue, orange, red, and gray color represent cytoplasm, cytoplasmic membrane, outer membrane, extracellular region, cell wall, periplasmic region and unknown subcellular localization respectively (iii) The third red ring in orange represents GC percentage. Peaks represent an increase in GC percentage (iv) fourth ring in blue color shows GC skew values for the whole genome. (v) The innermost red colored ring shows AT skew values calculated for the whole genome. The figure is generated by Circular Visualization for microbial genomes (CiVi).

Such metabolic models illustrates metabolism in a way which coincides with the known genetics and biochemistry of the organism. Furthermore these genome-scale metabolic network models can be computationally grilled by compiling them into mathematical models and then using a variety of constraint based modeling approaches to interpret metabolic networks.

7.3.1 Gene knockout analysis

Gene knockout studies have been a massive advancement in the biomedical and pharmaceutical sector. Gene knockout experiments help researchers to understand the role of a particular gene/reaction by searching a metabolic network for the genes that are critical to the production of biomass. Experimental validation of gene essentiality is expensive, and time-consuming [32] where in computational methods seemingly streamline the gap between rapidly mounting sequencing data and whole genome approaches to gene essentiality prediction [33-35].

Constraint-based modeling provides pragmatic solutions which are amenable to experimental testing by imposing certain constraints on biomass, growth etc. For identifying the genes which are indispensable to the bacterial growth, we used standard FBA and MOMA models implemented in OptFlux as outlined in the workflow chart in Figure 7.2. We identified 73 genes to be critical using gene knockout studies which are listed in Table 7.1. These genes rendered a reduction of greater than 90 percent in the total biomass yield during *in silico* knockout experiments using MOMA model and FBA model. In MOMA, mutant strain tries to minimize the adjustment of fluxes from wild type distribution, and thus the linear problem of FBA converts to an optimization problem with a quadratic objective function. The distribution pattern of these genes among various KEGG pathways is depicted in Figure 7.3.

7.3.2 Identification of essential non-homologous pathogen proteins

A BLASTP search of the proteins corresponding to these critical genes by homology search against human proteome rendered 65 genes as non-homologous to *H. sapiens* below e-value 0.05. Identifying drug targets which are absent from host organism is an important factor to assessing drug safety, as it excludes the possibility of any potential undesired adverse-effects which can arise upon drug administration. Adverse drug reactions add significantly to morbidity with a huge burden on hospitals globally [36]. These proteins were further filtered by performing a BLASTP search against experimentally predicted essential genes of *Salmonella enterica* serovar *Typhi*, *Salmonella enterica* serovar *Typhi* Ty2, *Salmonella enterica* serovar *Typhimurium* SL1344, *Salmonella enterica* subsp. *enterica* serovar *Typhimurium* str. 14028S, and *Salmonella typhimurium* LT2 in DEG. This search revealed that a total of 54 proteins were in alignment with experimentally verified essential genes from *Salmonella* spp.

7.3.3 Prioritization of drug targets

The essential gene set retrieved from this study was further prioritized through a multistep process which involved following criteria: a) molecular weight b) availability of experimental data on target crystal structure; c) subcellular localization; d) presence of transmembrane helices; e) druggability potential; and f) Assayability; to be reduced to a set of candidate drug targets. All the 54 proteins retrieved in previous step were having molecular weight <110 kDa which makes them amenable to isolation and experimental verification as small weight proteins are easier to purify [37].

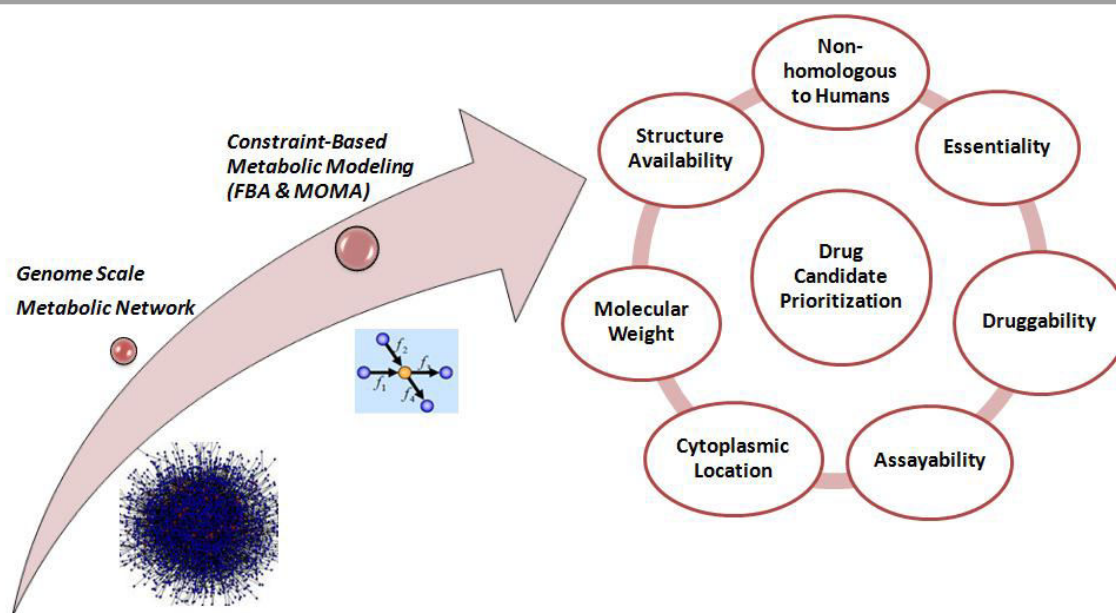


Figure 7.2: Schematic representation of the steps involved in target protein prioritization from genome scale metabolic network of *S. typhimurium* LT2.

In a search for 3D structural data, it was observed that crystal structure information was available for yadF, ribH, fabD, aroD, rfbG, rfbD, ppnK, and ribB proteins. For 9 proteins i.e. mraY, fabI, ribA, pgsA, folB, rfaL, dfp, cdsA, and psd no structural data was found to be existing. Rest of the proteins had structures solved computationally through homology modeling with varying sequence identities from 11-99% with the template and which were deposited in Modbase database. Even after a decade of *S. typhimurium* LT2 genome being sequenced out, structural information is typically lacking for the proteins encoded by this strain, which points towards a gap in structural characterization of pathogen proteins. Availability of target structure is an important factor which contributes largely to the drug identification, optimization and validation, thereby reducing the cost of high throughput experimental assays radically [38]. Typically majority of the drug targets were predicted to be of cytoplasmic origin as per PSORTb calculations. Subcellular localization of the drug target can shed some light on the protein function [20]. With combined predictions of PSORTb and TMHMM server, 6 proteins viz. mraY, cdsA, pgsA, plsC, rfaL were predicted to be membrane localized with one or more helices traversing the membrane, opening up new avenues to explore these proteins as vaccine candidates. All these results are summarized in Table 7.2.

Table 7.1: Characterization of essential therapeutic drug targets.

KEGG index	Gene Name	Uniprot Id	Protein Name	M. Wt (Da)	Length	PDB Id	Mod -base	% Identity
STM0045	ribF	Q7CR93	Riboflavin biosynthesis protein	34,271	312		Yes	34
STM0087	folA	Q8ZRW1	Dihydrofolate reductase	17,970	159		Yes	96
STM0123	murE	Q8ZRU7	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate--2,6-diaminopimelate ligase	53,287	495		Yes	91
STM0124	murF	Q8ZRU6	UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase	47,533	452		Yes	86
STM0125	mraY	Q8ZRU5	Phospho-N-acetylmuramoyl-pentapeptide-transferase	40,023	360		No	
STM0126	murD	Q8ZRU4	UDP-N-acetylmuramoylalanine--D-glutamate ligase	46,998	438		Yes	91
STM0128	murG	Q8ZRU3	UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol	37,863	355		Yes	92
			N-acetylglucosamine transferase					
STM0129	murC	Q8ZRU2	UDP-N-acetylmuramate--L-alanine ligase	53,438	491		Yes	69
STM0171	yadF	Q8ZRS0	Carbonic anhydrase	24,821	220	3QY1		
STM0183	folK	Q8ZRO9	7, 8-dihydro-6-hydroxymethylpterin-pyrophosphokinase, PPPK	17,742	159		Yes	83
STM0222	cdsA	Q7CR63	Phosphatidate cytidyltransferase	31,395	285			
STM0227	fabZ	P0A1H9	3-hydroxyacyl-[acyl-carrier-protein] dehydratase FabZ	16,999	151		Yes	56
STM0232	accA	P0A1C3	Acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha	35,344	319		Yes	12
STM0416	ribD	Q8ZRD4	Riboflavin biosynthesis protein RibD	39,730	367		Yes	89
STM0417	ribH	P66038	6,7-dimethyl-8-ribityllumazine synthase	16,008	156	3NQ4		
STM1067	fabA	P64105	3-hydroxydecanoyl-[acyl-carrier-protein] dehydratase	19,047	172		Yes	99
STM1194	fabD	O85140	Malonyl CoA-acyl carrier protein transacylase	32,405	309	3H0P		
STM1198	pabC	Q7CQR1	4-amino-4-deoxychorismate lyase	30,049	269		Yes	70
STM1200	tmk	P65246	Thymidylate kinase	23,724	213		Yes	13
STM1358	aroD	P58687	3-dehydroquinate dehydratase	27,325	252	3O1N		
STM1426	ribE	Q8ZPP1	Riboflavin synthase, alpha chain	23,392	213		Yes	91
STM1700	fabI	P16657	Enoyl-[acyl-carrier-protein] reductase [NADH] FabI	27,761	262		No	
STM1711	ribA	P66030	GTP cyclohydrolase-2	21,635	196		No	
STM1752	galU	Q7CQF2	UTP--glucose-1-phosphate uridylyltransferase	32,906	302		Yes	12
STM1824	pabB	P12680	Aminodeoxychorismate synthase component 1	50,979	454		Yes	76
STM1945	pgsA	Q7CQB9	CDP-diacetylglucosyl-glycerol-3-phosphate 3-phosphatidyltransferase	20,731	182		No	
STM2087	rfbV	P26401	Abequosyltransferase RfbV	38,579	333		Yes	12
STM2089	rfbJ	P0A1P4	CDP-abequose synthase	34,106	299		Yes	19
STM2090	rfbH	P26398	Lipopolysaccharide biosynthesis protein RfbH	48,104	437		Yes	14
STM2091	rfbG	P26397	CDP-glucose 4,6-dehydratase	41,022	359	1WV		
						G		
STM2092	rfbF	P26396	Glucose-1-phosphate cytidyltransferase	29,035	257		Yes	16
STM2093	rfbI	P26395	Protein RfbI	36,583	330		Yes	23
STM2094	rfbC	P26394	dTDP-4-dehydrothiamine 3,5-epimerase	20,663	183	1DZR		
STM2096	rfbD	P26392	dTDP-4-dehydrothiamine reductase	32,554	299	1N2S		

(continued)

Table 7.1 (CONTINUED)

KEGG index	Gene	Uniprot Id	Protein Name	M. Wt (Da)	Length	PDB Id	Mod -base	% Identity
STM2098	galF	P0A2K7	UTP-glucose-1-phosphate uridylyltransferase	32,958	297		Yes	22
STM2232	oafA	Q8ZNI3	Acetylation of the O-antigen (LPS)	69,648	609		No	
STM2365	folC	Q8ZNC0	Bifunctional protein FolC	45,481	422		Yes	30
STM2366	accD	Q7CQ41	Acetyl-coenzyme A carboxylase carboxyl transferase subunit beta	33,216	304		Yes	30
STM2378	fabB	Q7CQ36	3-oxoacyl-[acyl-carrier-protein] synthase 1	42,373	404		Yes	96
STM2384	aroC	P58729	Chorismate synthase	39,109	361		Yes	40
STM2652	pssA	Q7CQ03	CDP-diacylglycerol--serine O-phosphatidyltransferase	52,734	451		Yes	14
STM2683	ppnK	P65774	NAD kinase	32,584	292	2AN1		
STM3173	plsC	P0A257	1-acyl-sn-glycerol-3-phosphate acyltransferase	27,339	245		Yes	14
STM3195	ribB	P66032	3,4-dihydroxy-2-butanone 4-phosphate synthase	23,310	217	3LS6		
STM3205	uppP	P67388	Undecaprenyl-diphosphatase	29,792	273		No	
STM3206	folB		bifunctional dihydroneopterin aldolase/dihydroneopterin triphosphate 2'-epimerase	13,162	119		No	
STM3294	glmM	Q7CYP9	Phosphoglucosamine mutase	47,441	445		Yes	26
STM3295	folP	Q8ZLS8	Dihydropteroate synthase	30,484	282		Yes	92
STM3307	murA	P65454	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	44,741	419		Yes	93
STM3379	accB	Q7CPM1	AcetylCoA carboxylase, BCCP subunit	16,687	156		Yes	99
STM3469	pabA	P06193	Aminodeoxychorismate synthase component 2	20,801	187		Yes	16
STM3486	aroB	P77980	3-dehydroquinate synthase	38,710	362		Yes	13
STM3535	glgA	P05416	Glycogen synthase	52,946	477		Yes	11
STM3536	glgC	P05415	Glucose-1-phosphate adenylyltransferase	48,462	431		Yes	30
STM3713	rfaL	P26471	O-antigen ligase	46,034	404		No	
STM3725	coaD	Q8ZL48	Phosphopantetheine adenylyltransferase	17,956	159		Yes	93
STM3730	dfp		bifunctional phosphopantetheinylcysteine decarboxylase/phosphopantetheinate synthase	43,292	406		No	
STM3862	glmU	Q8ZKX0	Bifunctional protein GlmU	49,198	456		Yes	14
STM3926	wxE	Q9L6Q9	homologous to E. coli lipopolysaccharide biosynthesis protein	44,810	416		No	
STM4105	metF	P11003	5,10-methylenetetrahydrofolate reductase	33,173	296		No	
STM4137	murB	P37417	UDP-N-acetylenolpyruvoylglucosamine reductase	37,647	342		Yes	81
STM4139	coaA	Q9L9K3	Pantothenate kinase	36,236	316		Yes	95
STM4235	plsB	Q8ZKH9	Glycerol-3-phosphate acyltransferase	91,227	806		Yes	12
STM4348	psd	Q8ZKB1	Phosphatidylserine decarboxylase proenzyme	35,888	322		No	
STM4414	ppa	P65748	Inorganic pyrophosphatase	19,677	176		Yes	94

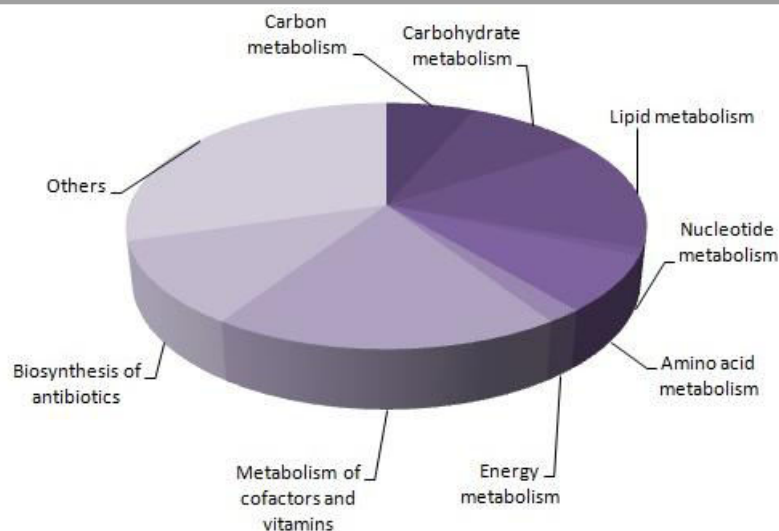


Figure 7.3: Distribution of the essential target proteins retrieved using FBA and MOMA approach in different pathways.

A BLASTP search of all the selected proteins was performed against the druggable proteins deposited in the DrugBank database to evaluate the druggability potential of the selected proteins. Druggability evaluated the probability of unearthing small molecules which alters the function of the drug targets to which they bind in a way that benefits the patient therapeutically [39]. We identified 47 proteins with a significantly higher similarity to the binders of these drug targeted proteins. Out of these, 14 proteins had FDA-approved drugs or nutraceuticals in the market against them, while for the rest of the proteins the drug binders were in the experimental stage only. In Table 7.2, for each drug target we have listed the identified drug types. Druggability potential is an important concern in drug target identification considering the fact that only 10-14% of the human genome is speculated to be druggable [40]. Information on assayability of each protein was referenced from BRENDA database. Assayability concept revolves around the availability of information on the biochemical or cellular assays for drug binding, inhibition/activation, and function. Assayability information is critical to support the screening programs for establishing suitable lead compounds. We checked the distribution pattern of proteins with regard to their involvement in various pathways before and after performing similarity search against DrugBank. The number of proteins after druggability potential evaluation was reduced by ~23% and there was a steep decrease in the number of proteins at every step.

Table 7.2: Similar binding partners of drugs available in Drug Bank inferred by homology search, druggability, assayability, α -helices and pathway information on prioritized drug target proteins of *S. typhimurium* LT2

Gene	Druggability	Drug Class	Essentiality	Location	TM-HMM	Assayability	Pathways Involved
ribF	✓	Experimental, Nutraceutical	✓	Cytoplasmic		✓	Riboflavin metabolism, Biosynthesis of secondary metabolites
folA	✓	Approved, Experimental, Nutraceutical	✓	Periplasmic		✓	Folate biosynthesis
murE	✓	Experimental	✓	Cytoplasmic		✓	Peptidoglycan biosynthesis, Lysine biosynthesis
murF	✓	Experimental	✓	Cytoplasmic		✓	Peptidoglycan biosynthesis, Vancomycin resistance, Lysine biosynthesis
mraY			✓	Cytoplasmic	10	✓	Peptidoglycan biosynthesis, Vancomycin resistance
				Membrane			
murD	✓	Experimental	✓	Unknown	1	✓	Peptidoglycan biosynthesis, D-Glutamine and D-glutamate metabolism
murG	✓	Experimental	✓	Cytoplasmic		✓	Peptidoglycan biosynthesis, Vancomycin resistance
				Membrane			
murC	✓	Experimental	✓	Cytoplasmic		✓	Peptidoglycan biosynthesis, D-Glutamine and D-glutamate metabolism
yadF			✓	Cytoplasmic			Nitrogen metabolism
cdsA			✓	Cytoplasmic	8	✓	Glycerophospholipid metabolism
				Membrane			
fabZ	✓	Experimental	✓	Cytoplasmic		✓	Fatty acid metabolism, Biotin metabolism
accA	✓	Approved, Experimental, Nutraceutical	✓	Cytoplasmic		✓	Fatty acid metabolism, Carbon metabolism, Biosynthesis of antibiotics, Microbial metabolism in diverse environments
ribD	✓	Experimental	✓	Cytoplasmic		✓	Riboflavin metabolism, Biosynthesis of secondary metabolites
ribH	✓	Experimental	✓	Cytoplasmic		✓	Riboflavin metabolism, Biosynthesis of secondary metabolites
fabA	✓	Experimental	✓	Cytoplasmic		✓	Fatty acid metabolism
fabD	✓	Experimental	✓	Cytoplasmic		✓	Fatty acid metabolism
pabC	✓	Experimental	✓	Cytoplasmic		✓	Folate biosynthesis
tmk	✓	Experimental	✓	Cytoplasmic		✓	Pyrimidine metabolism
aroD	✓	Experimental	✓	Cytoplasmic		✓	Phenylalanine, tyrosine and tryptophan biosynthesis, Biosynthesis of secondary metabolites, Biosynthesis of antibiotics
ribE	✓	Approved, Nutraceutical	✓	Cytoplasmic			Riboflavin metabolism, Biosynthesis of secondary metabolites
fabI		Approved, Experimental, Nutraceutical	✓	Cytoplasmic		✓	Fatty acid biosynthesis, Fatty acid metabolism, Biotin metabolism
				Membrane			
ribA		Experimental, Nutraceutical	✓	Cytoplasmic		✓	Riboflavin metabolism, Biosynthesis of secondary metabolites
galU	✓	Experimental, Nutraceutical	✓	Cytoplasmic			Amino sugar and nucleotide sugar metabolism, Biosynthesis of antibiotics
pabB	✓	Experimental	✓	Cytoplasmic		✓	Folate biosynthesis
pgsA	✓	Experimental	✓	Cytoplasmic	4	✓	Glycerophospholipid metabolism
				Membrane			
ribV			✓	Cytoplasmic			
ribJ	✓	Experimental	✓	Cytoplasmic		✓	Amino sugar and nucleotide sugar metabolism
ribH	✓	Experimental	✓	Cytoplasmic			Amino sugar and nucleotide sugar metabolism
ribG	✓	Experimental	✓	Cytoplasmic		✓	Amino sugar and nucleotide sugar metabolism

(continued)

Table 7.2 (CONTINUED)

Gene	Druggability	Drug Class	Essentiality	Location	TM-HMM	Assay-ability	Pathways Involved
rfbF	✓	Approved, Experimental	✓	Cytoplasmic			Amino sugar and nucleotide sugar metabolism
rfbI			✓	Cytoplasmic			Amino sugar and nucleotide sugar metabolism
rfbD	✓	Approved, Experimental, Nutraceutical	✓	Cytoplasmic		✓	Polyketide sugar unit biosynthesis, Streptomycin biosynthesis
galF	✓	Experimental, Nutraceutical	✓	Cytoplasmic			Amino sugar and nucleotide sugar metabolism, Biosynthesis of antibiotics
folC	✓	Approved, Experimental, Nutraceutical	✓	Cytoplasmic		✓	Folate biosynthesis
accD	✓	Approved, Experimental, Nutraceutical	✓	Cytoplasmic		✓	Fatty acid metabolism, Carbon metabolism, Biosynthesis of antibiotics, Microbial metabolism in diverse environments
fabB	✓	Approved, Experimental	✓	Cytoplasmic		✓	Fatty acid biosynthesis, Fatty acid metabolism, Biotin metabolism
aroC	✓	Approved, Experimental	✓	Cytoplasmic		✓	Phenylalanine, tyrosine and tryptophan biosynthesis, Biosynthesis of secondary metabolites, Biosynthesis of antibiotics
pssA			✓	Cytoplasmic		✓	Glycine, serine and threonine metabolism, Glycerophospholipid metabolism, Biosynthesis of secondary metabolites
ppnK			✓	Cytoplasmic		✓	Nicotinate and nicotinamide metabolism
plsC			✓	Cytoplasmic Membrane	I		Glycerophospholipid metabolism, Biosynthesis of secondary metabolites
ribB			✓	Cytoplasmic		✓	Riboflavin metabolism, Biosynthesis of secondary metabolites
folB	✓	Experimental	✓	Cytoplasmic		✓	Folate biosynthesis
glmM	✓	Experimental	✓	Cytoplasmic		✓	Amino sugar and nucleotide sugar metabolism, Biosynthesis of antibiotics
murA	✓	Approved, Experimental	✓	Cytoplasmic		✓	Amino sugar and nucleotide sugar metabolism, Peptidoglycan biosynthesis
accB	✓	Approved, Nutraceutical	✓	Cytoplasmic			Fatty acid metabolism, Carbon metabolism, Biosynthesis of antibiotics, Microbial metabolism in diverse environments
rfal			✓	Cytoplasmic Membrane	10		Lipopolysaccharide biosynthesis
coaD	✓	Experimental, Nutraceutical	✓	Cytoplasmic		✓	Pantothenate and CoA biosynthesis
dfp	✓	Approved, Experimental	✓	Cytoplasmic			Pantothenate and CoA biosynthesis
glmU	✓	Experimental, Nutraceutical	✓	Cytoplasmic		✓	Amino sugar and nucleotide sugar metabolism, Biosynthesis of antibiotics
murB	✓	Approved, Experimental	✓	Cytoplasmic		✓	Amino sugar and nucleotide sugar metabolism, Peptidoglycan biosynthesis
coaA	✓	Experimental, Nutraceutical	✓	Cytoplasmic		✓	Pantothenate and CoA biosynthesis
plsB	✓	Experimental, Nutraceutical	✓	Cytoplasmic Membrane		✓	Glycerophospholipid metabolism, Biosynthesis of secondary metabolites
psd	✓	Approved, Nutraceutical	✓	Cytoplasmic Membrane		✓	Glycerophospholipid metabolism, Biosynthesis of secondary metabolites
ppa	✓	Experimental	✓	Cytoplasmic		✓	Oxidative phosphorylation

The numbers of proteins which are present at each stage are summarized in Figure 7.4. The distribution pattern of various genes based on their involvement in pathways also shifted. After searching against DrugBank, the number of proteins in pathways such as Metabolism of cofactors and vitamins, Lipid metabolism, Carbohydrate metabolism, Biosynthesis of secondary metabolites, Glycan biosynthesis and metabolism reduced while for Biosynthesis of antibiotics, Carbon metabolism, Nucleotide metabolism, Amino acid metabolism, Metabolism of terpenoids and polyketides was more or less the same (Figure 7.5). After applying all the filtration criteria, we selected proteins involved in L-rhamnose biosynthesis, peptidoglycan biosynthesis, fatty acid biosynthesis, folate biosynthesis as our serving set of top priority therapeutic drug candidates.

PPIs are intrinsic to various aspects of virtually every biological process and PPI networks are important from perspectives of system level understanding of the cellular processes. Protein expression is a dynamic process which responds to a multitude of stimuli. Analyzing interacting partners of a protein can open up new avenues for cross-species predictions and also the network embodies information on the functional modularity and interconnectivity in the cell [27]. Our protein interaction analysis results showed that selected drug targets were involved in crucial cross talks between themselves and there were direct interactions between these proteins.

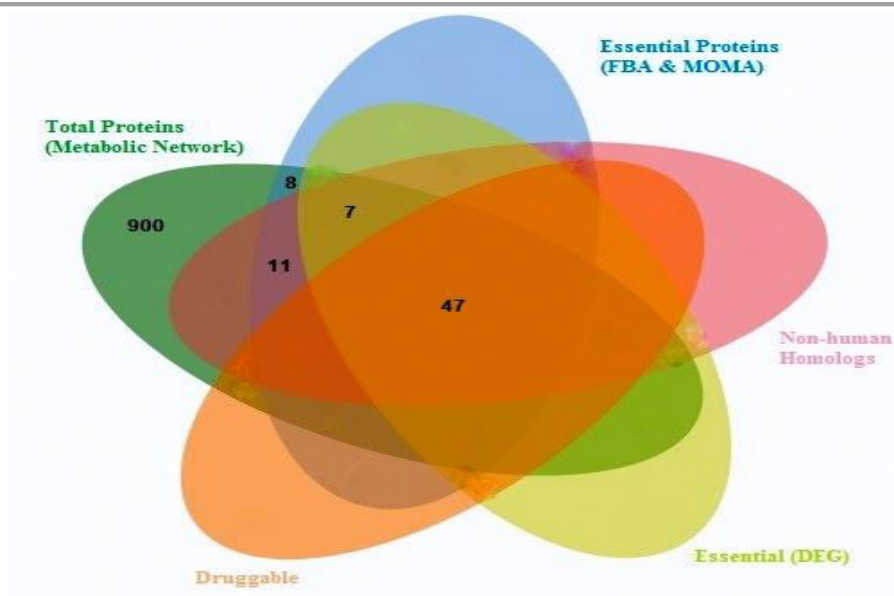


Figure 7.4: Representation of 47 prioritized target proteins: The proteins characterized as i) proteins with a role in metabolic network ii) essential proteins using combined predictions of

FBA and MOMA approach iii) Non-homologous to humans iv) experimentally predicted as essential from prediction of DEG v) druggable, are shown using different colors in the Venn diagram. Proteins satisfying a particular parameter are shown in the corresponding category of the Venn diagram. Three proteins were prioritized as drug targets. The image has been generated using Jvenn.

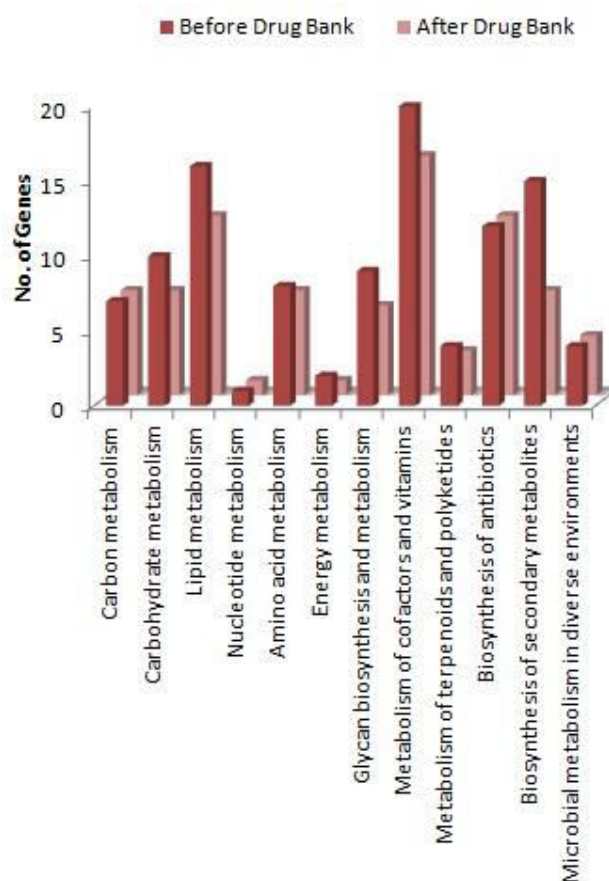


Figure 7.5: Distribution of essential non-homologous proteins of *S. typhimurium* LT2 in different pathways before and after drug bank search for druggability potential evaluation.

The topological analysis of the network as shown in Figure 7.6, revealed that folC and floP enzymes were interacting with pabA, pabB, pabC, folA, and mur ligases which were mapped to pathways such as peptidoglycan biosynthesis, streptomycin synthesis, amino acid and nucleotide synthesis. Involvement in these pathways signifies their association with pathogenesis mechanism and bacterial survival. For the selected drug targets, their structures were predicted using RaptorX, where no information was available on crystal structure or homology modeled

structure. Structures of proteins pabA, glmM, glmU, mraY, rfbF, and rfbH predicted through homology modeling are shown in Figure 7.7.

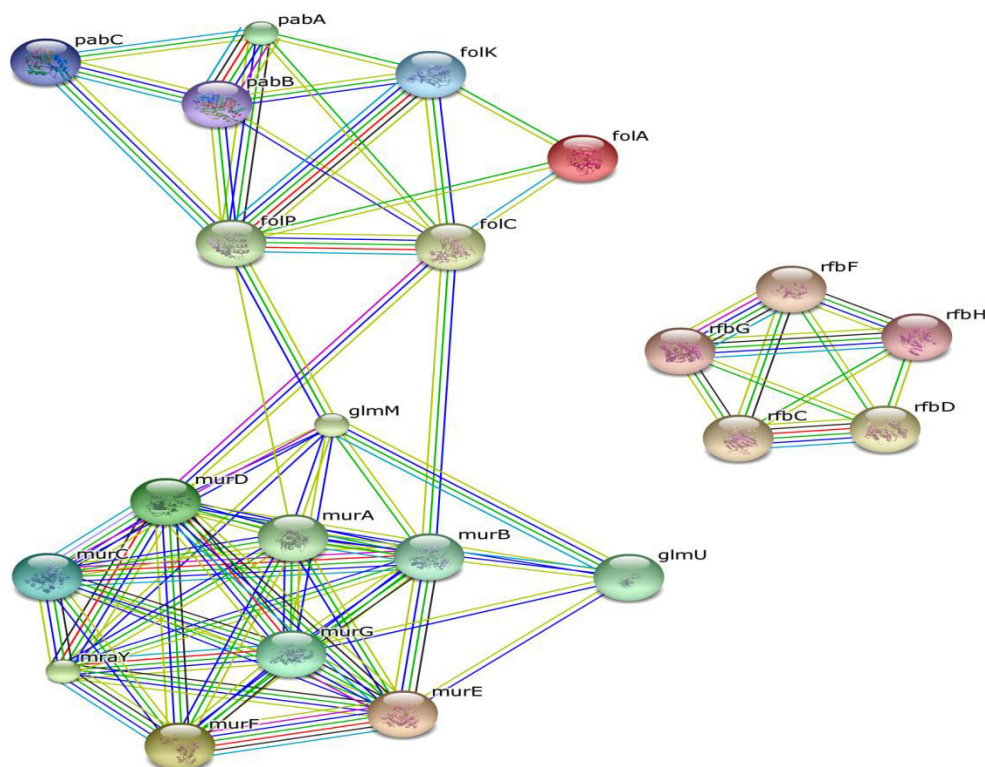


Figure 7.6: Protein–protein interactions among the prioritized proteins. Proteins involved in peptidoglycan synthesis, folate synthesis show extensive linkages among them. Proteins involved in L-rhamnose synthesis are rich in interactions among themselves.

Bacterial cell wall deserves special attention owing to its fundamentally different nature from eukaryotic counterpart and to presence of various virulence determinants encoded in bacterial membrane. Complex carbohydrate structures and lipids are essential components of bacterial cell wall [41]. L-rhamnose, a 6-deoxyhexose is one such building block of bacterial cell wall, absent from mammals. L-rhamnose is utilized by gram negative bacteria by incorporating it into O-antigen of various serotypes [42]. The *rfb* gene cluster of *S. typhimurium* LT2 encoding *rfbC* (dTDP-4-dehydrorhamnose 3,5-epimerase), *rfbD* (dTDP-4-dehydrorhamnose reductase), *rfbF* (Glucose-1-phosphate cytidyltransferase), *rfbG* (CDP-glucose 4,6-dehydratase), *rfbH* (Lipopolysaccharide biosynthesis protein RfbH) is essential for expression of O-antigen; which

is highly polymorphic and predominantly determines the pathogenesis in *Salmonella* spp. and *E. coli* [43]. rfbC catalyzes the epimerization of the C3' and C5' positions of dTDP-6-deoxy-D-xylo-4-hexulose, forming dTDP-6-deoxy-L-lyxo-4-hexulose and rfbD catalyzes the final step of L-rhamnose biosynthesis pathway by reducing dTDP-6-deoxy-L-lyxo-4-hexulose to dTDP-L-rhamnose as shown in Figure 7.8 [44]. rfbD has been speculated to be an attractive drug target due to its probability to have specific drug pockets against rhamnose like compounds [42]. rfbF is involved in tyvelose biosynthesis by catalyzing the transfer of CMP moiety to glucose 1-phosphate. rfbG and rfbH are both involved in the LPS O-antigen biosynthesis, which is a unique component of bacterial outer membrane, toxic to animals [45].

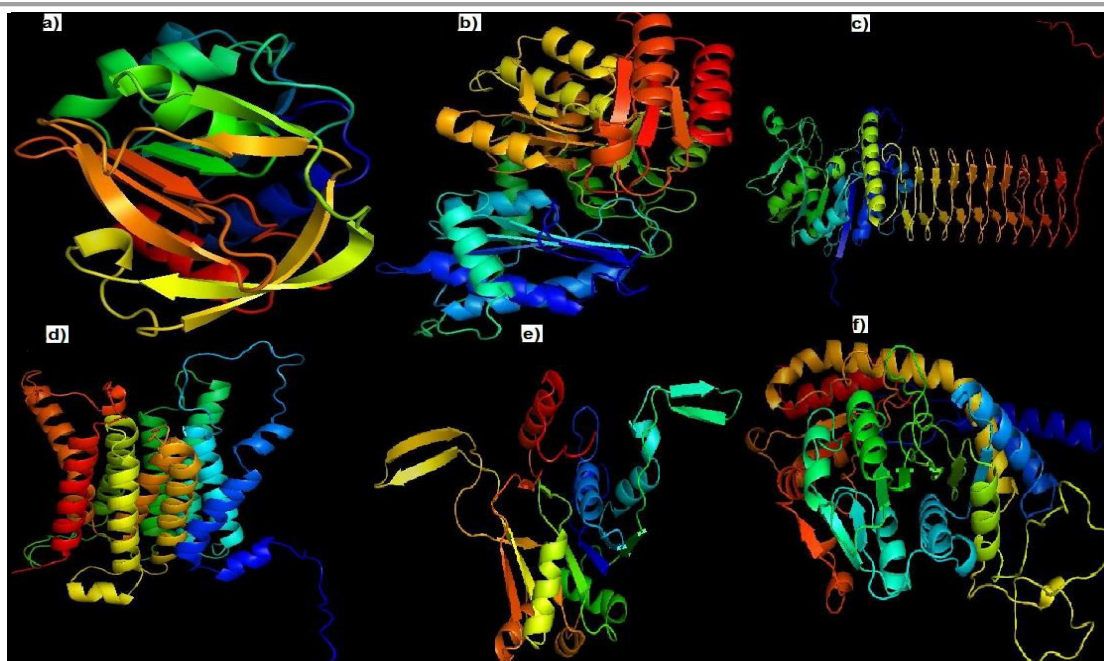


Figure 7.7: 3D structures of the prioritized proteins a) pabA, b) glmM, c) glmU, d) mraY, e) rfbF, and f) rfbH for which no crystal structure information was available based on novel nonlinear context-specific alignment.

Recently peptidoglycan biosynthesis pathway has been extensively subjected to drug targeting. As the enzymes involved in peptidoglycan biosynthesis are indispensable to bacterial survival and do not have any orthologs in humans [46], which has directed researchers to an intensified research for discovering new drug targets against bacterial pathogens. Peptidoglycan biosynthesis is a multistage process. Stage I occurring in cytoplasm involves the formation of

monomeric building block, N-acetylglucosamine-N-acetylmuramyl pentapeptide through action of ten cytoplasmic enzymes. glmM (Phosphoglucosamine mutase) and glmU (Bifunctional protein GlmU) are important stage I enzymes which synthesize UDP-N-acetylglucosamine (UNAG) [46]. UDP-N-acetylmuramic acid (UNAM) synthesis from UNAG involves murA (UDP-N-acetylglucosamine 1-carboxyvinyltransferase) and murB (UDP-N-acetylenolpyruvoylglucosamine reductase) catalyze important reactions in the peptidoglycan precursor synthesis. murA transfers an enolpyruvyl group from phosphoenolpyruvate (PEP) to UNAG to form UDP-N-acetylglucosamine enolpyruvate (UNAGEP) followed by a murB-catalyzed NADPH dependent reduction of UNAGEP to UNAM. murB has previously been reviewed to be essential in *E. coli* [47]. murC to murF genes located in a mra cluster assemble the peptide moiety of the monomeric unit of peptidoglycan, thus called as Mur ligases. murG and mraY are stage II enzymes which synthesize an intermediate monomer unit of the peptidoglycan polymer [48]. Majority of antibiotics in clinical use today target later steps of peptidoglycan synthesis [49]. All these enzymes are highly conserved sharing common structural motifs and indispensable to bacterial cell survival [50]. Thus they embrace great opportunities to be exploited as therapeutic drug targets.

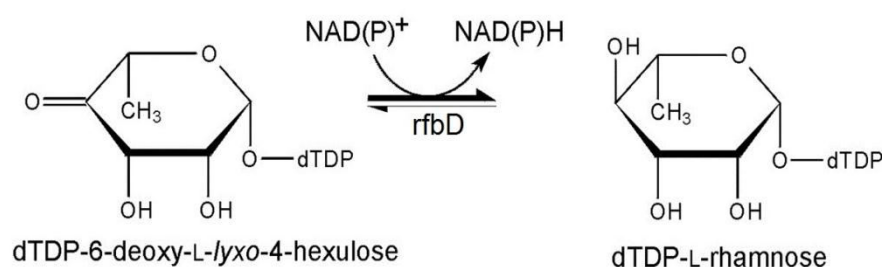


Figure 7.8: Final Step catalyzed by rfbD in dTDP-L-rhamnose biosynthesis pathway viz. reduction of dTDP-6-deoxy-L-lyxo-4-hexulose to dTDP-L-rhamnose.⁴⁴

Folate is a critical precursor for the synthesis of a variety of cellular components such as purine, pyrimidine and amino acids by catalyzing one-carbon transfer reactions. Unlike higher eukaryotes, in bacteria folate must be synthesized *de novo* through folate biosynthesis pathway. folA (Dihydrofolate reductase), folC (Bifunctional protein FolC), folK (7, 8-dihydro-6-hydroxymethylpterin-pyrophosphokinase, PPPK), folP (Dihydropteroate synthase),

pabA (Aminodeoxychorismate synthase component 2), pabB (Aminodeoxychorismate synthase component 1), and pabC (4-amino-4-deoxychorismate lyase) are important enzymes of folate biosynthesis pathway. folA is a critical node in folate synthesis as it links folate synthesis and utilization to tetrahydrofolate production (THF) as shown in Figure 7.9. folA effectiveness as an antimicrobial agent is widely acknowledged [51]. Pterin component upstream of folA in folate synthesis pathway is also synthesized *de novo*. Pterin synthesis is controlled by enzymes such as folC, folK, folP rendering them attractive drug targets as eukaryotes produce tetrahydrobiopterin through a different pathway and pterin has important functions in nerve signaling and maintenance of homeostasis of brain chemistry [52]. Folate biosynthesis pathway has significant untapped potential to be exploited as drug targets as the mutations in the enzymes involved in this pathway are reported to produce non-viable phenotypes [53,54] and moreover this pathway is absent in higher eukaryotes as they salvage membrane associated folate proteins [55].

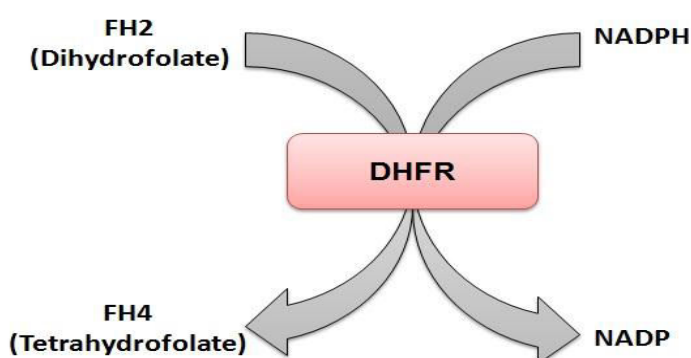


Figure 7.9: A critical node in folate synthesis, Dihydrofolate reductase, links folate synthesis to tetrahydrofolate production (THF) utilization by reducing dihydrofolate to tetrahydrofolate.

Although fatty acid synthesis pathways are virtually similar between mammals and bacteria, but different profile and arrangements of active sites in these enzymes makes it profoundly trailed pathway for drug targets [56]. fabI (Enoyl-[acyl-carrier-protein] reductase [NADH] FabI), one of the most intensively explored target is no longer considered as viable due to resistance to fabI-specific drugs in organism carrying fabI gene and failure to achieve broad spectrum antibacterial activity [57]. fabA (3-hydroxydecanoyl-[acyl-carrier-protein] dehydratase) and fabZ (3-hydroxyacyl-[acyl-carrier-protein] dehydratase FabZ) are essential for unsaturated fatty acid synthesis, but fabZ is considered as the most attractive drug target as it is uniformly expressed in

bacterial system while *fabA* distribution is limited [56]. Riboflavin is a central player in cell physiology as it serves as precursor to many redox reactions as a component of flavin adenine dinucleotide (FAD) and Flavin mononucleotide (FMN). *ribD* (Riboflavin biosynthesis protein RibD), *ribH* (6,7-dimethyl-8-ribityllumazine synthase), *ribE* (Riboflavin synthase, alpha chain) were predicted to be good drug targets supplemented with the fact that riboflavin is synthesized *de novo* in bacteria, a pathway absent from mammals [58]. The ubiquitous enzyme *ppa* (Inorganic pyrophosphatase) catalyzing inorganic pyrophosphate (PP_i) hydrolysis to *ortho*-phosphate (P_i), is essential for bacterial survival [59] and has important role in energy and lipid metabolism [60] making it an attractive drug target.

Thus, summing up we can conclude that the approach that we have utilized is statistically robust as many of the identified candidate drug targets have already been exploited as drug targets and many of them have already existing antibiotics targeted towards them; which is apparent from a search performed against Drug bank database. We have identified a variety of other metabolites which have not yet been targeted. Majority of the metabolites identified as potential drug targets are involved in cell wall synthesis.

Conclusion

Genome-scale metabolic networks can provide valuable insights into the complex molecular mechanisms governing cellular aspects of pathogen metabolism in pathophysiological state. To resolve gaps in our understanding of pathogen metabolism and host-pathogen relationship, single target-centric approaches have extended themselves into system-level studies. In this study, constraint-based analysis methods have been applied to large scale metabolic network reconstruction model of *S. typhimurium* to identify candidate drug targets. Identifying drug targets tailored according to a particular pathogen's need will result in narrow-spectrum antibiotics against these drug targets which will ultimately enable to surmount the menace of multi-drug resistance. We believe, this general therapeutic approach offers a way to identify unique metabolites with a potential to be explored as drug target rapidly and this systems-level approach can be applied to other microbes as well to combat the global threat of multi-drug resistant bacteria.

References

1. WHO. Diarrhoeal disease. (World Health Organization, 2013)
2. Rabsch W, Andrews HL, Kingsley RA *et al.* Salmonella enterica Serotype Typhimurium and Its Host-Adapted Variants. *Infection and Immunity*, 70(5), 2249-2255 (2002).
3. Graham SM, Molyneux EM, Walsh AL, Cheesbrough JS, Molyneux ME, Hart CA. Nontyphoidal Salmonella infections of children in tropical Africa. *The Pediatric Infectious Disease Journal*, 19(12), 1189-1196 (2000).
4. Gordon MA, Graham SM, Walsh AL *et al.* Epidemics of Invasive Salmonella enterica Serovar Enteritidis and S. enterica Serovar Typhimurium Infection Associated with Multidrug Resistance among Adults and Children in Malawi. *Clinical Infectious Diseases*, 46(7), 963-969 (2008).
5. DiMarzio M, Shariat N, Kariyawasam S, Barrangou R, Dudley EG. Antibiotic resistance in Salmonella Typhimurium associates with CRISPR sequence type. *Antimicrobial Agents and Chemotherapy*, (2013).
6. Seyfarth AM, Wegener HC, Frimodt-Moller N. Antimicrobial resistance in Salmonella enterica subsp. enterica serovar typhimurium from humans and production animals. *Journal of Antimicrobial Chemotherapy*, 40(1), 67-75 (1997).
7. Su LH, Chiu CH, Chu C, Ou JT. Antimicrobial Resistance in Nontyphoid Salmonella Serotypes: A Global Challenge. *Clinical Infectious Diseases*, 39(4), 546-551 (2004).
8. Blake DP, Hillman K, Fenlon DR, Low JC. Transfer of antibiotic resistance between commensal and pathogenic members of the Enterobacteriaceae under ileal conditions. *Journal of Applied Microbiology*, 95(3), 428-436 (2003).
9. Livermore DM. The need for new antibiotics. *Clinical Microbiology and Infection*, 10, 1-9 (2004).
10. Zavlanos MM, Julius AA. Robust flux balance analysis of metabolic network. *American Control Conference*, 2915-2920 (2011).
11. Ma H, Goryanin I. Human metabolic network reconstruction and its impact on drug discovery and development. *Drug Discovery Today*, 13(9-10), 402-408 (2008).
12. Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Micro*, 2(11), 886-897 (2004).
13. Jamshidi N, Palsson BA. Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. *BMC Systems Biology*, 1, 26-26 (2007).
14. Oberhardt MA, Puchalka J, Fryer KE, Martins dos Santos VtAP, Papin JA. Genome-Scale Metabolic Network Analysis of the Opportunistic Pathogen Pseudomonas aeruginosa PAO1. *Journal of Bacteriology*, 190(8), 2790-2803 (2008).
15. Lee DS, Burd H, Liu J *et al.* Comparative Genome-Scale Metabolic Reconstruction and Flux Balance Analysis of Multiple Staphylococcus aureus Genomes Identify Novel Antimicrobial Drug Targets. *Journal of Bacteriology*, 191(12), 4015-4024 (2009).

16. Kim HU, Kim TY, Lee SY. Genome-scale metabolic network analysis and drug targeting of multi-drug resistant pathogen *Acinetobacter baumannii* AYE. *Molecular BioSystems*, 6(2), 339-348 (2010).
17. Raghunathan A, Reed J, Shin S, Palsson B, Daefer S. Constraint-based analysis of metabolic capacity of *Salmonella typhimurium* during host-pathogen interaction. *BMC Systems Biology*, 3, 38-38 (2009).
18. Rocha I, Maia P, Evangelista P *et al.* OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Systems Biology*, 4(1), 1-12 (2010).
19. Zhang R, Ou HY, Zhang CT. DEG: a database of essential genes. *Nucleic Acids Research*, 32(suppl 1), D271-D272 (2004).
20. Yu NY, Wagner JR, Laird MR *et al.* PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26(13), 1608-1615 (2010).
21. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1. *Journal of Molecular Biology*, 305(3), 567-580 (2001).
22. Berman HM, Westbrook J, Feng Z *et al.* The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235-242 (2000).
23. Pieper U, Webb BM, Dong GQ *et al.* ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research*, 42(D1), D336-D346 (2014).
24. Haas J, Roth S, Arnold K *et al.* The Protein Model Portal-a comprehensive resource for protein structure and model information. *Database*, 2013 (2013).
25. Kallberg M, Margaryan G, Wang S, Ma J, Xu J. RaptorX server: A Resource for Template-Based Protein Structure Modeling. In: *Protein Structure Prediction*. Kihara, D (Ed. (Springer New York, New York, NY, 2014) 17-27.
26. Law V, Knox C, Djoumbou Y *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, 42(D1), D1091-D1097 (2014).
27. Franceschini A, Szklarczyk D, Frankild S *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(Database issue), D808-D815 (2013).
28. McClelland M, Sanderson KE, Spieth J *et al.* Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature*, 413(6858), 852-856 (2001).
29. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784), 299-304 (2000).
30. Blanc-Potard A-Ba, Solomon F, Kayser J, Groisman EA. The SPI-3 Pathogenicity Island of *Salmonella enterica*. *Journal of Bacteriology*, 181(3), 998-1004 (1999).
31. Parkhill J, Dougan G, James KD *et al.* Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature*, 413(6858), 848-852 (2001).
32. Butt AM, Nasrullah I, Tahir S, Tong Y. Comparative Genomics Analysis of *Mycobacterium ulcerans* for the Identification of Putative Essential Genes and Therapeutic Candidates. *PLoS ONE*, 7(8), e43080 (2012).

-
33. Agüero F, Al-Lazikani B, Aslett M *et al.* Genomic-scale prioritization of drug targets: the TDR Targets database. *Nat Rev Drug Discov*, 7(11), 900-907 (2008).
 34. Doyle MA, Gasser RB, Woodcroft BJ, Hall RS, Ralph SA. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics*, 11(1), 1-14 (2010).
 35. Volker C, Brown JR. Bioinformatics and the Discovery of Novel Anti-Microbial Targets. *Current Drug Targets - Infectious Disorders*, 2(4), p279 (2002).
 36. Davies EC, Green CF, Taylor S, Williamson PR, Mottram DR, Pirmohamed M. Adverse Drug Reactions in Hospital In-Patients: A Prospective Analysis of 3695 Patient-Episodes. *PLoS ONE*, 4(2), e4439 (2009).
 37. Duffield M, Cooper I, McAlister E, Bayliss M, Ford D, Oyston P. Predicting conserved essential genes in bacteria: in silico identification of putative drug targets. *Molecular BioSystems*, 6(12), 2482-2489 (2010).
 38. Marianne AG. Protein Structure Prediction in Structure-Based Ligand Design and Virtual Screening. *Combinatorial Chemistry & High Throughput Screening*, 12(10), 940-960 (2009).
 39. Edfeldt FNB, Folmer RHA, Breeze AL. Fragment screening to predict druggability (ligandability) and lead discovery success. *Drug Discovery Today*, 16(7-8), 284-287 (2011).
 40. Hopkins AL, Groom CR. The druggable genome. *Nat Rev Drug Discov*, 1(9), 727-730 (2002).
 41. Silhavy TJ, Kahne D, Walker S. The Bacterial Cell Envelope. *Cold Spring Harbor Perspectives in Biology*, 2(5), a000414 (2010).
 42. Blankenfheldt W, Kerr ID, Giraud M-F *et al.* Variation on a Theme of SDR. *Structure*, 10(6), 773-786 (2002).
 43. Jiang XM, Neal B, Santiago F, Lee SJ, Romana LK, Reeves PR. Structure and sequence of the rfb (O antigen) gene cluster of *Salmonella* serovar typhimurium (strain LT2). *Molecular Microbiology*, 5(3), 695-713 (1991).
 44. Graninger M, Nidetzky B, Heinrichs DE, Whitfield C, Messner P. Characterization of dTDP-4-dehydrorhamnose 3,5-Epimerase and dTDP-4-dehydrorhamnose Reductase, required for dTDP-L-rhamnose Biosynthesis in *Salmonella enterica* Serovar Typhimurium LT2. *The Journal of Biological Chemistry*, 274, 25069-25077 (1999).
 45. Salton MRJ. Structure and Function of Bacterial Cell Membranes. *Annual Review of Microbiology*, 21(1), 417-442 (1967).
 46. Gautam A, Vyas R, Tewari R. Peptidoglycan biosynthesis machinery: A rich source of drug targets. *Critical Reviews in Biotechnology*, 31(4), 295-336 (2011).
 47. Pucci MJ, Discotto LF, Dougherty TJ. Cloning and identification of the *Escherichia coli* murB DNA sequence, which encodes UDP-N-acetylenolpyruvoylglucosamine reductase. *Journal of Bacteriology*, 174(5), 1690-1693 (1992).
 48. Bouhss A, Trunkfield AE, Bugg TDH, Mengin-Lecreux D. The biosynthesis of peptidoglycan lipid-linked intermediates. *FEMS Microbiology Reviews*, 32(2), 208-233 (2008).
 49. El Zoeiby A, Sanschagrin F, Levesque RC. Structure and function of the Mur enzymes: development of novel inhibitors. *Molecular Microbiology*, 47(1), 1-12 (2003).

50. Bouhss A, Mengin-Lecreulx D, Blanot D, van Heijenoort J, Parquet C. Invariant Amino Acids in the Mur Peptide Synthetases of Bacterial Peptidoglycan Synthesis and Their Modification by Site-Directed Mutagenesis in the UDP-MurNAc:l-Alanine Ligase from *Escherichia coli*. *Biochemistry*, 36(39), 11556-11563 (1997).
51. Swarbrick J, Iliades W, Simpson J, Macreadie I. Folate biosynthesis- Reappraisal of Old and Novel Targets in the Search for New Antimicrobials. *The Open Enzyme Inhibition Journal*, 1, 12-33 (2008).
52. Haruki H, Hovius R, Pedersen MG, Johnsson K. Tetrahydrobiopterin Biosynthesis as a Potential Target of the Kynurenine Pathway Metabolite Xanthurenic Acid. *Journal of Biological Chemistry*, 291(2), 652-657 (2016).
53. Fermer C, Swedberg G. Adaptation to sulfonamide resistance in *Neisseria meningitidis* may have required compensatory changes to retain enzyme function: kinetic analysis of dihydropteroate synthases from *N. meningitidis* expressed in a knockout mutant of *Escherichia coli*. *Journal of Bacteriology*, 179(3), 831-837 (1997).
54. Pyne C, Bognar AL. Replacement of the *folC* gene, encoding folylpolyglutamate synthetase-dihydrofolate synthetase in *Escherichia coli*, with genes mutagenized in vitro. *Journal of Bacteriology*, 174(6), 1750-1759 (1992).
55. Henderson GB, Huennekens FM. [42] Membrane-Associated folate transport proteins. In: *Methods in Enzymology*. (Academic Press, 1986) 260-269.
56. Campbell JW, Cronan JE. Bacterial Fatty Acid Biosynthesis: Targets for Antibacterial Drug Discovery. *Annual Review of Microbiology*, 55(1), 305-332 (2001).
57. Zhang Y-M, White SW, Rock CO. Inhibiting Bacterial Fatty Acid Synthesis. *Journal of Biological Chemistry*, 281(26), 17541-17544 (2006).
58. Long Q, Ji L, Wang H, Xie J. Riboflavin Biosynthetic and Regulatory Factors as Potential Novel Anti-Infective Drug Targets. *Chemical Biology & Drug Design*, 75(4), 339-347 (2010).
59. Chen J, Brevet A, Fromant M *et al.* Pyrophosphatase is essential for growth of *Escherichia coli*. *Journal of Bacteriology*, 172(10), 5686-5689 (1990).
60. Jukka K. H. Biological Role of Inorganic Pyrophosphate. (Springer US, 2001) 250.

Closing Remarks: Conclusion and Future Perspectives

Conclusion

Diarrheal diseases, a highly common infection, remain a leading cause of childhood mortality and morbidity worldwide. Diarrheal diseases are a serious impediment to cognitive development, childhood growth and survival, predominantly in the developing countries. Continuous replenishment of lost electrolytes is a critical treatment for diarrheal management which is accomplished by intake of commercially prepared Oral Rehydration Solution (ORS) in conjunction with Zinc supplementation. Most cases of diarrhea are self-limiting, occasionally requiring antibiotic treatment in case of TD, persistent and invasive diarrhea and *C. difficile* infection. The Global Burden of Disease (GBD) study, 2013 estimated that diarrheal mortality rate was significantly reduced by almost 31.1% from 1.8 million in 2000 to 1.3 million in 2013, however the proportional mortality rate still is soaring with an estimated 1.7 billion diarrheal incidences annually. A wide variety of etiological agents are implicated to be associated with diarrheal diseases including viruses, bacteria, protozoa, and helminthes. Due to increasing antibiotic resistance among diarrheal pathogens, limited surveillance of emerging antibiotic resistant pathogens, irrational use of antibiotics and lack of efforts devoted to drug development, tackling the problem of the global mass killer diarrhea seems a goal which cannot be achieved in the near future.

We have used diarrheal diseases as our case study and built a systematic computational pipeline to study mechanisms of antibiotic resistance and propose effective solutions to address the knowledge gaps in containment of antibiotic resistance. In the first part of thesis, we have designed a database, DBDiaSNP which is a comprehensive repository of mutations and resistance genes among various diarrheal pathogens and hosts to advance breakthroughs that will find applications from development of sequence based diagnostic tools to drug discovery. It contains information about 946 mutations and 326 resistance genes compiled from published literature. DBDiaSNP is the first antibiotic resistance database dedicated to the diarrheal pathogens covering mutations and resistance genes that have clinical relevance from a broad range of pathogens and hosts. For future translational research involving integrative biology and global health, the database offers veritable potentials, particularly for developing countries and worldwide monitoring and personalized effective treatment of pathogens associated with

diarrhea. The database is accessible on the public domain at <http://www.juit.ac.in/attachments/dbdiasnp/>.

In the second part, the mutations deposited in DBDiaSNP database were mapped to structural level using molecular modeling and molecular dynamics simulations to gain atomistic details into the mechanism of antibiotic resistance. Several mutations in the quinolone resistance determining region (QRDR) of the gene *gyrA*, an attractive drug target for quinolones, are associated with increased quinolone resistance *in vivo*. Fully atomistic explicit-water solvated molecular dynamics simulations of wild-type and mutant forms of *gyrA* in ETEC and *C. jejuni*, complexed with ciprofloxacin reveal that significant conformational changes are not observed upon introducing mutations which is reflected by stable RMSD over the course of simulation across all mutant and wild type structures; nonetheless these mutations drastically alter *gyrA* residue interaction network and the overall pattern of global dominant motions in major distinctive domains of N-terminal regions of *gyrA*.

In the last part of the thesis, we have explored the innovative approaches based on systems biology and exploited the power of integrated omics approaches to predict potential vaccine candidates and drug targets against ETEC, *Campylobacter* spp. and *Salmonella* spp. Integrated comparative genomics and immunoinformatics approach for proteome scale identification of peptide vaccine candidates were utilized. To predict potential vaccine candidates, the proteins shared between pathogenic strains but lacking in commensal strain were subjected to immunoinformatics analysis based on physicochemical and immunogenic properties like population coverage in diarrhea endemic regions; presentation by major MHC-I supertypes and efficient binding in MHC groove. In *Salmonella*, we performed flux balance analysis and minimization of metabolic adjustment studies of genome scale reconstruction model of the pathogen to identify large number of metabolites with a potential to be utilized as therapeutic drug targets. Metabolites involved in L-rhamnose biosynthesis, peptidoglycan biosynthesis, fatty acid biosynthesis, and folate biosynthesis pathways were prioritized as candidate drug targets. In *Campylobacter*, comparative metabolic pathway analysis approach integrated with subtractive genomics was utilized to facilitate rapid identification of drug targets. Majority of the predicted drug targets had indispensable role in cell wall synthesis and energy metabolism. Such computational approaches can yield insights into novel protective antigens which could better

guide the fate of future medicine ultimately making the vaccine and drug development a more robust and competent task.

Future Perspectives

With a significant reduction in the sequencing cost, the availability of data on mutation frequency through sequencing projects is expected in near future. Due to accessibility to mutation frequency data, it would be feasible to develop SNPs based diagnostic methods which would identify mutations associated with a particular disease more rapidly and with high accuracy; and to comprehend the consequences of regional differences in drug resistance. Mapping of mutations at the structural level reveal the molecular mechanisms underlying mutations associated antibiotic resistance and consequently can be further exploited to aid in design of more potent antibacterial agents with high ligand efficacy for treating drug resistant bacterial infections. Virtual screening of inhibitors against the predicted drug targets proteins may lead to discovery of novel therapeutic compounds which will be pathogen specific and have minimal toxic effect on the host. Predicted epitopes open up new avenues for experimental validation using model organisms which will eventually accelerate the development of successful vaccines to control diarrheal disease burden which has plagued human lives for years. These computational approaches can be further extended to understanding host-pathogen interactions which would further ease the treatment of infectious diseases by shedding light on the complex interplay between host and pathogen.

We conclude by stating that the omics approaches described in this thesis hopefully lend to a host of new drug targets and vaccine candidates against diarrheal pathogens. The database DBDiaSNP will prove to be useful to researchers and clinicians working in the diarrheal diseases. Indeed, the progress in understanding antibiotic resistance from computational perspective is quite electrifying; and there will be many more breakthroughs in the future.

ANNEXURE

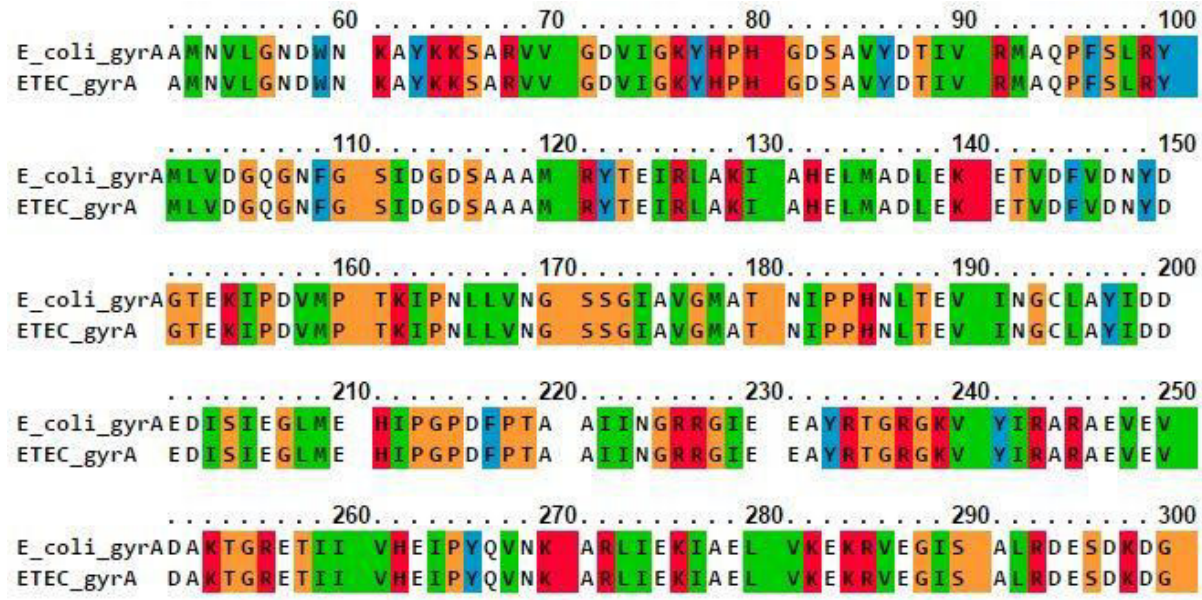


Figure 1: A snippet of sequence alignment for ETEC gyrA and template structure of gyrA from *E. coli* showing 91.2% identity.

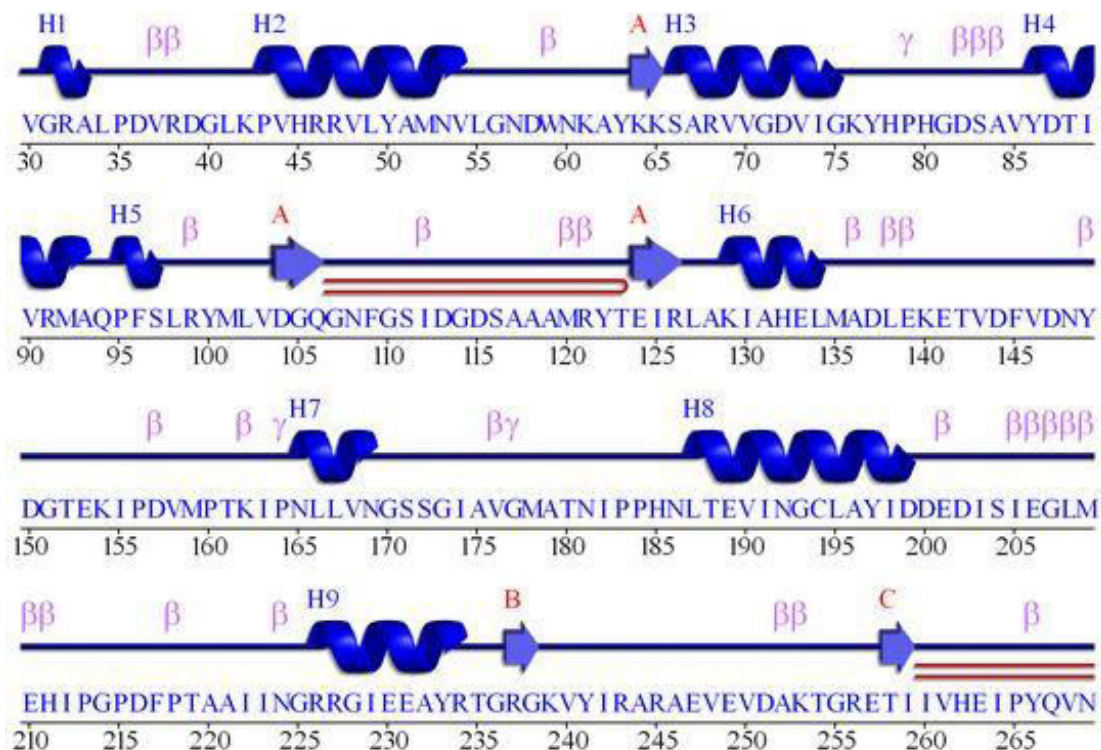


Figure 2: A snippet of the secondary structure elements assignment in the corresponding sequence of the modeled protein structure of ETEC gyrA.

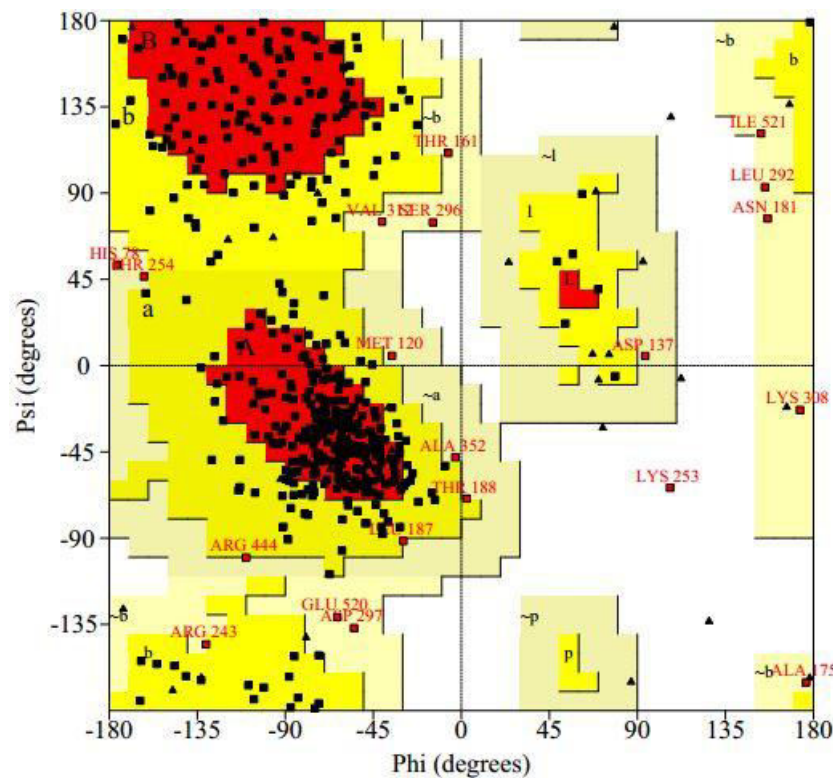


Figure 3: Ramachandran plot of the *E. coli* gyrA based ETEC gyrA model generated with PROCHECK.

30	VGRALPDV	RDGLKPVHRRVLYAMNVL	GNDWNKAYKKSARVVG	DVI GKYHHPHGDS	AVYDTI VRMAQPF	SLR	99	<i>E. coli</i>
33	I GRALPDAR	DGLKPVHRRILYAMQNDEAKSRTDFVKSARI	VGAVI GRYHHPGDT	AVYDALVRMAQDF	SMR		102	<i>C. jejuni</i>
100	YMLVDGGG	NFGSI DGDSAAAMRYTEI	RLAKI AHELMADLEKETVDFVDNYDGTEKI	PDVMPTKI	PNLLVN		169	<i>E. coli</i>
103	YPSITGGG	NFGSI DGDSAAAMRYTEAKMSKLSHELLKDI	DKDITVDFVPNYDGSESEPDVLP	SRVPNLLN			172	<i>C. jejuni</i>
170	GSSGI	AVGMATNIPPHNLTEVINGCLAYIDDEDISIEGLMEHIPGPDFPTAAI	INGRRGIEEAYRTGRGK				239	<i>E. coli</i>
173	GSSGI	AVGMATNIPPHSLNELIDGLLYLLDNKDASLEEI	MQFI KGPDFPTGGI	IYGKKGI	IEAYRTGRGR		242	<i>C. jejuni</i>
240	VYI	RARAEVEVET- - - - IIVHEIPYQVNKARLIEKI	AELVKEKRV	EGISALRDESDKDG	MRIVIE- - - -		306	<i>E. coli</i>
243	VKVRAKTHI	EKKTNKDVIVIDELPYQTNKARLIEQI	AELVKEKRV	EGISEVRDES	NKEGIRVVI	ELKREA	312	<i>C. jejuni</i>
313	-	GEVVLNNLYSQTQLQVSFGINMVALHHGQPKIMNLKDI	IAAFVRHRRREV	VTRRTIFELRKARDRAHILE			381	<i>E. coli</i>
313	M	SEIVLNNLFKSTTMESTFGVIMLAIHNKEPKIFSLLELLNLF	LTHRKIVIRRTIFELQKARARAHILE				382	<i>C. jejuni</i>
382	ALAVALANI	DPIIELIRHAPTPAEAKTALVAN	PWQLGNVAAML	EDAAARPEWLEPEFGVRDGL	YYLTEQQA		455	<i>E. coli</i>
383	GLKI	ALDNI DEVI ALIKNSSDNNTARDSLVAK	- - - - -	- - - - -	FGLSELQA		422	<i>C. jejuni</i>
456	QAI	LDLRLQKLTGLEHEKLLDEYKELLDQIAELLRI	LGSADRLMEVIREEL	ELVREQFGDKRRTEIT			522	<i>E. coli</i>
423	NAI	LDMKLGRLTGLEREKIENELAELMKEI	ARLEEILKSETLLENLI	RDELKEIRSKFDVPRI	IQIE		489	<i>C. jejuni</i>

Figure 4: Sequence alignment for *C. jejuni* gyrA and template structure of gyrA from *E. coli*.

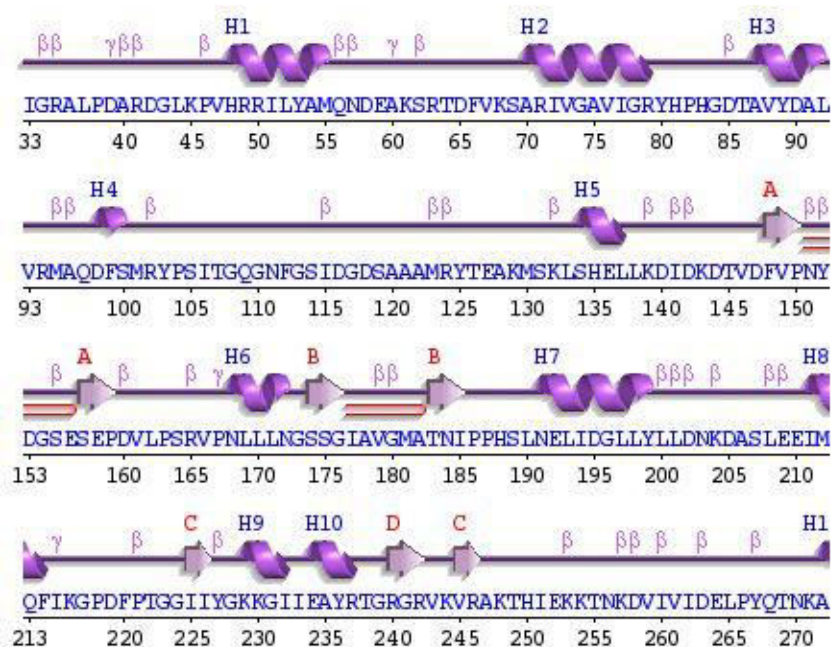


Figure 5: A snapshot of the secondary structure assignment in the corresponding sequence of the modeled protein structure of *C. jejuni* gyrA.

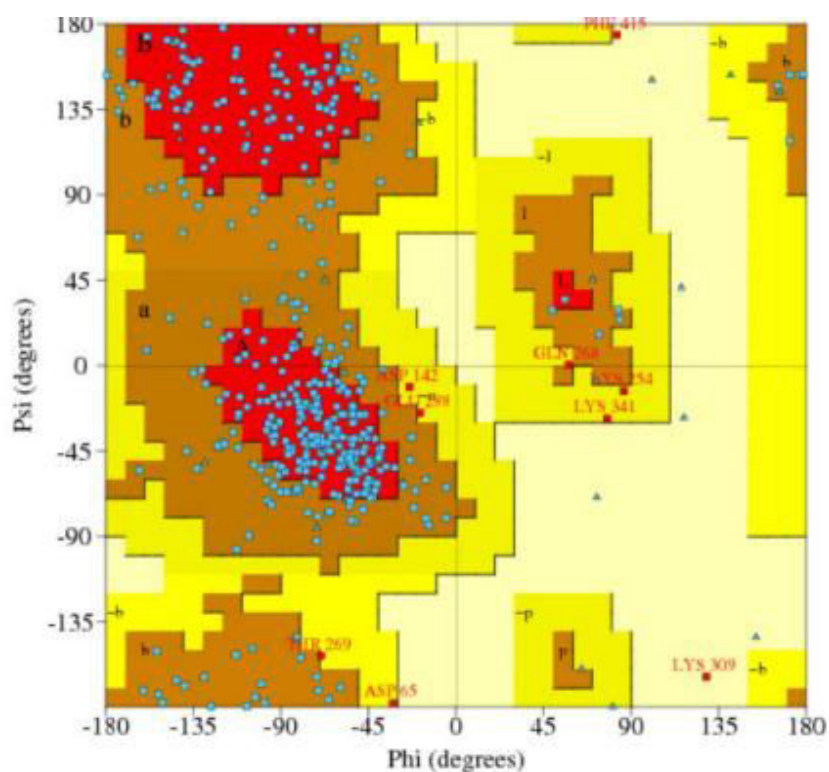


Figure 6: Ramachandran plot of the modeled structure of gyrA from *C. jejuni* generated using PROCHECK.

PUBLICATIONS AND CONFERENCES

PUBLICATIONS

1. **Mehla K.**, Ramana J. DBDiaSNP: An Open-Source Knowledgebase of Genetic Polymorphisms and Resistance Genes in Diarrheal Pathogens. *OMICS: A Journal of Integrative Biology* 2015; 19(6):354-60. [IF:2.896]
2. **Mehla K.**, Ramana J. Novel Drug Targets for Foodborne Pathogen *Campylobacter jejuni*: An Integrated Subtractive Genomics and Comparative Metabolic Pathway Study. *OMICS: A Journal of Integrative Biology* 2015; 19(7):393-406. [IF: 2.896]
3. **Mehla K.**, Ramana J. Structural signature of Ser83Leu and Asp87Asn mutations in DNA gyrase from Enterotoxigenic *Escherichia coli* and impact on quinolone resistance. *Gene* 2016; 576(1):28–35. [IF:2.319]
4. **Mehla K.**, Ramana J. Identification of epitope-based peptide vaccine candidates against enterotoxigenic *Escherichia coli*: a comparative genomics and immunoinformatics approach. *Mol Biosyst.* 2016; 12(3):890-901 [IF:3.2]
5. **Mehla K.**, Ramana J. Travelers' Diarrhea-associated enterotoxigenic *Escherichia coli* gyrA mutants and quinolone antibiotic affinity: a molecular dynamics simulation and residue interaction network analysis. *OMICS: A Journal of Integrative Biology* 2016; 20(11):635-644. [I.F.: 2.896]
6. **Mehla K.**, Ramana J. Surface Proteome Mining for Identification of Potential Vaccine Candidates against *Campylobacter jejuni*: An In silico approach. *Funct Integr Genomics.* 2017; 17(1):27-37. [I.F.: 2.265]
7. **Mehla K.**, Ramana J. Tapping into *Salmonella typhimurium* LT2 genome in a quest to explore its therapeutic arsenal: a metabolic network modeling approach. *Biomed Pharmacother.* 2017; 86:57-66. [I.F.: 2.326]
8. **Mehla K.**, Ramana J. Molecular Dynamics Simulations of quinolone resistance associated T86I and P104S mutations in *C. jejuni* gyrA: Unraveling structural repercussions. *Microb Drug Resist.* Accepted. [I.F.: 2.529]

CONFERENCE PRESENTATIONS

1. **Mehla K.**, Ramana J. A Docking Study of Relationship between Point Mutations and Ciprofloxacin Resistance in *Campylobacter jejuni*. **Oral Presentation** at IC LIFE 2014: International Conference on Life Sciences, Informatics, Food and Environment, August 29-30, 2014 at IIIT, Noida
2. **Mehla K.**, Ramana J. Constraint based analysis of *Clostridium difficile* metabolic network: FBA and MOMA approach for drug target identification. **Oral Presentation** at International Conference on Advances in Biomedical Engineering, Cancer Biology, Stem Cells, Bioinformatics and Applied Biotechnology (ABECBAB-2016) , May 21, 2016 at JNU, Delhi.