Genomics of some pathogenic food bacteria and molecular modeling of their important toxins and their interactions

THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF
PHILOSOPHY (SCIENCE) OF THE
UNIVERSITY OF NORTH BENGAL

by

Uttam Kumar Mondal

Supervisor
Dr. Asim K. Bothra

Co-supervisor
Dr. Arnab Sen

UTTAM KUMAR MONDAL
DEPARTMENT OF CHEMISTRY
RAIGANJ COLLEGE (UNIVERSITY COLLEGE)
RAIGANJ, UTTAR DINAJPUR
2010

DEDICATED TO MY PARENTS

DECLARATION

I hereby declare that the research work embodied in this thesis entitled *"Genomics of some pathogenic food bacteria and molecular modeling of their important toxins and their interactions"* has been carried out in the Department of Chemistry of Raiganj College (University College), under the joint supervision of Dr. Asim K. Bothra, Reader, Raiganj College, (University College) & Dr. Arnab Sen, Reader, Department of Botany, University of North Bengal.

To the best of my knowledge and belief, this thesis or any part of it has not been submitted for the award of PhD or any other degree or diploma at any University or Institute

Date: 27.12.10
Place: Raiganj

Uttam Kr. Mondal
(Uttam Kumar Mondal)

उत्तर बंग
विश्वविद्यालय

UNIVERSITY OF NORTH BENGAL
RajaRammohumpur•Siliguri 734013•West Bengal•India

CERTIFICATE

We certify that the thesis entitled, "Genomics of some pathogenic food bacteria and molecular modeling of their important toxins and their interactions" submitted by Mr. Uttam Kumar Mondal for the award of PhD degree from Department of Chemistry of Raiganj College (University College), embodies the record of the original investigation carried by him under our supervision. He has been duly registered and the thesis presented is worthy of being considered for the award of Doctor of Philosophy (Science) degree in Chemistry. The work has not been submitted for any degree of this or any other university and is in accordance with the rules and regulations of the University of North Bengal.

Asim Bothra

(Asim K Bothra)
Supervisor
Reader
Department of Chemistry
Raiganj College(University College)

Date: 27.12.10

READER
Dept. Of Chemistry
Raiganj College (University College)

(Arnab Sen)
Co-Supervisor
Reader
Department of Botany
University of North Bengal

Date: 27.12.10

Dr. Arnab Sen
Reader in Botany
University of North Bengal

ACKNOWLEDGEMENT

Apart from personal efforts and steadfastness to work, constant inspiration and encouragement given by a number of individuals served as the driving force that enabled me to submit this thesis in the present form. Inspiration, guidance, direction, co-operation, all came in abundance and it seems almost an impossible task for me to acknowledge the same in adequate terms.

First of all, I take this opportunity to express my deep sense of gratitude & sincere thanks to my guide (Supervisor), Dr. Asim K. Bothra, Reader, Department of Chemistry, Raiganj College (University College) for his mature, able & invaluable guidance & persistent encouragement. No words can express my sincere and deep sense of reverence for him.

I owe heartfelt thanks to Dr. Arnab Sen, Reader, Department of Botany, North Bengal University for the constant support and guidance throughout the work. I am extremely thankful to him for his invaluable suggestions and advices.

I am also thankful to Dr. P. K. Mondal Ex Principal, and Dr. Dilip De Sarker Teacher-in-Charge, Raiganj College (University College) for providing me facilities as required from time to time & active interest in the progress of this work. My thanks are due to all other faculty members of the Chemistry Department & all the members of library of Raiganj College (University College) for their help and moral support during my work.

I am thankful to all the member of Chemistry and Botany Department of University of North Bengal for their inspiration.

Words are insufficient to express my gratefulness and indebtedness to Prof T.C. Ghosh, Bose institute, Kolkata for his suggestion & encouragement during my work and Dr. S.K. Gupta, Dr. Surajit Basak, Sujata Roy for providing research paper.

I feel lacuna of words to express my gratefulness and indebtedness to all my colleagues who have directly or indirectly helped me to present this work in present form.

My sincere thanks to my friends and lab mates Saubashya Sur, Shyamal Sharma, Biswajit Das, Bhaskar Bagchi, Dipanjan Sarkar, Abhick Chatterjee, Dr. Apurba Kanti Deb, Subarna Thakur, Dr. Biplop Sarkar, Dr. Arjun Ghosh, Dr. Shaonli Das, Tanmay Chowdhury, Arindam Dutta and my student Debjani Chakraborty for offering me assistance.

My special thanks to Ayan Pal for providing biological backup in the smooth completion of this work.

Thanks are not enough for Sujajta Bothra who allow me in her home every time as and when required for computational work and sometimes directly help me in computational work.

I remember to my intimate friend Dipak K. Mahato for his inspiration and Asishda & Nayan for molar support.

All my family members contributed significantly to bring this day in my life. My parents remained a constant source of strength throughout my educational career and later in sustaining my academic interest for higher learning. I thank my father and mother for providing me the moral support and resources to finish my work.

I am also thankful to many other persons whose name may not figure out here for their help & support.

In the end, I am thankful to the Almighty for blessing me to complete this work successfully.

Dated: 27.12.10
Place: Raiganj.

Uttam Kr. Mondal
(Uttam Kumar Mondal)

Summary

Different types of pathogenic bacteria cause food poisoning. In most cases, food poisoning are caused by *Staphylococcus aureus, Salmonella sp, Clostridium perfringens, Campylobacter sp, Listeria monocytogenes., Vibrio parahaemolyticus, Bacillus cereus,* and entero-pathogenic *Escherichia coli* which produce a number of toxins. These bacteria are commonly found in raw foods. Since food-pathogenic bacteria are often present in many foods, knowing the characteristics of such bacteria is essential for effective diagnosis and control.

The commonality in microbial virulence mechanisms and the occurrences of similar resistance systems in animals and plants point out that all these mechanisms have an ancient and intertwined history. It is quite evident that susceptibility or resistance to disease involves subtle and highly specific exchanges of molecular signals between pathogens and their hosts and a clear-cut understanding of these mechanisms can provide newer approaches to diagnose and control diseases. The genomic islands and operons are considered as the units where groups of genes are transcribed together and whose products contribute to specific function. One of the typical examples of genomic islands is the pathogenic island (PAIs), which is present in pathogenic bacteria that form the principal molecular component responsible for the development of a specific disease. Codon usage study provides information of use of different codon in a genome, as it is often seen that all codons are not used evenly. A detailed and accurate analysis of codon usage is an essential prerequisite to our understanding of how and why divergent patterns of codon choice evolved. Genomics has a great potential in the study of food pathogens showed the relationship between the predicted level of gene expression based on codon usage, actual microarray expression values and gene function at the genomic level in *S. pneumonie.*

Although a lot of work has been performed and is going on in codon usage of different microorganisms but very little work performed on codon usage of food pathogenic bacteria.

To understand the mode of action of toxin it is very much necessary to known their three dimensional structure and motional properties. Crystal structures of several toxins have been solved which enlighten us about their structure. But the structures of many toxins are yet to be solved. Molecular modeling &molecular dynamics of some toxin have been done.

In this work firstly we would try to characterize the pathogenicity island and toxic genes by statistical analysis, secondly knowledge based model will be built to get three-dimensional structure of toxins. This thesis contains total eight chapters. First and Second chapter describe introduction and review of this work. Third chapter describe codon usage patterns of five complete genomes of *Salmonella*, predict highly expressed genes, examine horizontally transferred pathogenicity-related genes to detect their presence in the strains, and scrutinize the nature of highly expressed genes to infer upon their lifestyle where Protein coding genes, ribosomal protein genes, and pathogenicity-related genes were analysed with Codon W and CAI (codon adaptation index) Calculator. Fourth Chapter describe Bioinformatic study of Pathogenicity related genes of three species of Helicobactor where Protein coding genes, ribosomal protein genes, and pathogenicity-related genes were analysed with Codon W and E-CAI (codon adaptation index) server. In the fifth chapter describe Characterization of pathogenic genes through condensed matrix method, case study through bacterial zeta toxin. In this study, zeta toxin nucleotide sequences of some pathogenic and non-pathogenic Bacteria were used for phylogenetic analysis. The uniqueness of this method is that it does not employ sequence alignment of complete nucleotide sequence of the corresponding gene.

In the chapter six describe Molecular Dynamics Simulation Receptor-Binding C-Terminal Domain from *Clostridium difficile* Toxin A to understand the motional properties and mode of action of the receptor-binding C-terminal domain of *C. difficile*, molecular dynamics simulation of *C. difficile* toxin A in aqueous environment was carried out using the software, GROMACS. From the time evolution, Root Mean Square Deviation (RMSD), Root Mean Square Fluctuation (RMSF) and Radius of gyration (Rg), it was found that the toxin was relatively inflexible. Principal Component Analysis (PCA) was also performed for better understanding of motional properties in reduced dimension. Analysis of binding site reveals that Ala51, Ala58, Ile59 and Tyr93 have very low fluctuation. All these observations help us to understand the mechanism of pathogenesis related with toxin A of *C. difficile.*

In the chapter seven describe Comparison in motional properties of *Staphylococcus aureus* exfoliative toxins A and B as revealed by their MD simulation to understand the motional properties and mode of action of *Staphylococcus aureus* exfoliative toxins A and B, molecular dynamics simulation are carried out using the software GROMACS. From the time evolution RMSD, RMSF and Radius of gyration, it is found that the toxin A and B are not much flexible and it is also indicated by principal component analysis. Pro192 and val183 is key residue towards the activity of toxin A and toxin B respectively. It is seen dihedral angle psi of Pro192 of toxin A is free to rotate without involvement of much energy and its becomes active after a conformational triggering of this dihedral but Val 183, the corresponding residue of toxin B, shows very less conformational freedom.

In the last chapter describe Homology modeling and MD simulation of the CdtB of *Helicobacter hepaticus* ATCC 51449 where the 3D model of the CdtB of *H. hepaticus*ATCC is constructed by MODELLER 9v4 program using the templates CdtB from *Actinobacillus actinomycetemcomitans*. The model is validated by PROCHECK, ProSa, CASTp server, ProFunc server etc. After that molecular dynamics simulation is performed using GROMACS and the resulting trajectory is analyzed. Homology modeling can produce high-quality structural models when the target and template are closely related, which has inspired the formation of a structural genomics consortium dedicated to the production of representative experimental structures for all classes of protein folds (Williamson AR, 2000). Like other methods of structure prediction, current practice in homology modeling is assessed in a biannual large-scale experiment known as the Critical Assessment of Techniques for Protein Structure Prediction, or CASP.

The wealth of information obtained from the genome projects needs to be mined. As newer and newer toxins genes are discovered and sequenced novel insights are being gained. Bioinformatics in combination with metagenomics as well as metaproteomics approaches has the potential to give a more detailed picture that underlies pathogenicity as well as diseases. In absence of crysllographic or NMR structure Homology modeling will enlighten us about three dimensional structure and molecular dynamics simulation opening newer possibilities for exploring the molecular mechanism and activity of that toxin. At the end of the day scientific perception will continue to play a vital role in creating models that clarify the functions of pathogens in improved manner.

Publication

Sl. No.	Journal Name	Article Title	Status
1	*Indian Journal of Medical Microbiology*	**Bioinformatic study of Pathogenicity related genes of three species of *Helicobacter***	**Communicated**
2	***International Journal of Biotechnology & Biochemistry***	**Characterization of pathogenic genes through condensed matrix method, case studt through bacterial zeta toxin.**	**Under review**
3	***International Journal of interrogative Biology***	**Homology modeling and MD simulation of the Cytolethal distending toxin B gene of *Helicobacter hepaticus* ATCC 51449**	**Published (2010) 10(1): 35-40**
4	***International Journal of interrogative Biology***	**Similarities and differences in motional properties of *Staphylococcus aureus* exfoliative toxins A and B as revealed by their Molecular dynamics simulation**	**Published (2009) 8(1): 19-24**
5	***The IUP Journal of Biotechnology***	**Molecular Dynamics Simulation Receptor-Binding C-Terminal Domain from *Clostridium difficile* Toxin A**	**Published (2009) 3(3): 22-37**
6	***Indian Journal of Medical Microbiology,***	**Comparative analysis of codon usage patterns and identification of predicted highly expressed genes in five *Salmonella* genomes**	**Published (2008) 26(4): 313-21**
7	***Journal of Biomolecular Structure & Dynamics,***	**Nucleotide Triplet Based Molecular Phylogeny of Class I and Class II Aminoacyl t-RNA Synthetase in Three Domain of Life Process: Bacteria, Archaea, and Eukarya**	**Published (2008) 26(3): 321-328**

CONTENTS

Page No.

Chapter I

Introduction

1.1: Pathogen & pathogenicity & toxins

A pathogen, commonly known as germ is an infectious biological agent that causes disease to its host. In 1999, Casadevall and Pirofski defined pathogen as a microbe capable of causing host damage. This definition can encompass classical pathogens and opportunistic pathogens and host damage can result from either direct microbial action or the host immune response. The term pathogenicity is defined as the capacity of a microbe to cause damage in a host. Early views of pathogenicity and virulence were primarily pathogen centered and were based on the assumption that these characteristics were intrinsic properties of microorganisms, although it was recognized that pathogenicity was neither invariant nor absolute (Casadevall & Pirofski 1999).

Many pathogenic organisms grow in food. Food borne illnesses are a widespread public health problem all around the world. At present it accounts for about 20 million cases annually in the world. Developing countries bear the brunt of the problem due to the presence of a wide range of food-borne diseases. In India an estimated 4,00,000 children below five years age die each year due to diarrhoea which is a common form of food borne disease. Recent studies reveal that food-borne diseases are a serious health hazard and important cause of morbidity and mortality in developing countries. Most cases go unreported and scientific investigations are rarely feasible. Studies carried out during 1995-2005 showed that the incidence of food-borne disease outbreaks were due to microorganisms like *Salmonella* and *Campylobacter jejuni* (Sudershan *et al.,* 2009). Different types of pathogenic bacteria cause food borne disease, namely *Staphylococcus aureus, Salmonella sp., Clostridium perfringens, Campylobacter sp., Listeria monocytogenes, Vibrio parahaemolyticus, Bacillus cereus,* and entero-pathogenic *Escherichia coli* which produce a number of toxins. These bacteria are commonly found in raw foods. In terms of mortality rates the major bacterial food-borne pathogens are *Salmonella* (non-typhoidal), *Listeria monocytogenes*, *Campylobacter*, and enterohaemorrhagic *Escherichia coli (*Mead *et al.*, 1999).

The pattern of foodborne disease has changed substantially in industrialized countries in recent decades. Outbreaks are more likely to be far reaching, and some are even global in scale because of widespread food distribution methods and changes in ways of food preparation. Further changes in the incidence of foodborne disease and the pattern of food-related illness can be anticipated from global warming. As a result of changed conditions in food production and better laboratory detection techniques, new foodborne pathogens continue to be identified. In particular, we are now faced with the emergence of antimicrobial drug–resistant bacteria and a number of viruses not previously recognized (Hall *et al.,* 2005). The incidence of human diseases caused by food-borne pathogens, such as *Salmonella* serotypes, *Staphylococcus aureus*, *Campylobacter jejuni* and Campylobacter coli, enterotoxigenic and enteroinvasive *Escherichia coli, Clostridium peljringens,* and *Bacillus cereus,* has not decreased. In fact, changes in lifestyle have increased the opportunities for transmission of the pathogenic bacteria through foods. In addition, emergence of new pathogens (*Yersinia enterocolitica, Listeria monocytogenes, E. coli* 0157:H7, *Aeromonas spp., Plesiomonas spp.*) or the emergence of specific subtypes of a species associated with a specific food (*Salmonella* serotype *Enteritidis* in eggs) has caused the redirection of resources from control programs for other well-known food-borne pathogens (Swaminathan & Feng 1994).The genetic makeup of bacterial genomes is subject to rapid and dramatic change through a variety of processes collectively referred to as "horizontal gene transfer". Recent evidence has shown that horizontal gene transfer plays a principal part in the molecular evolution of novel bacterial pathogens (Ochman & Moran 2001; Wren 2000; Ziebuhr *et al.,* 1999). Horizontal gene transfer refers to the incorporation of genetic elements transferred from a donor organism directly into the genome of the recipient organism, where they form genomic islands. Genomic islands may contain large blocks of virulence determinants (adhesins, invasins, toxins, protein secretion systems, antibiotic resistance mechanisms, etc) and thus are referred to as pathogenicity islands. Pathogenicity islands were first described in pathogenic species of *E coli*, but have since been found in the genomes of numerous bacterial pathogens of humans, animals, and plants (*Salmonella*, *Vibrio*, *Shigella*, *Yersinia*, *Listeria*, *S aureus*, etc) (Garcia et al., 1999; Lindsay *et al.,* 1998). Pathogenicity islands are believed to have been acquired as a block by horizontal gene transfer because of their G+C content is significantly different from that of the genomes of the host micro-organism and they are often flanked by direct repeats. PI are associated with tRNA genes they are associated with integrase determinants and other mobility loci and they exhibit genetic instability (Hacker & Kaper 2000). The determination and analysis of the complete genomic sequences of several important bacterial pathogens has led to the revelation that horizontal gene transfer may be much more extensive than previously thought of (Wren 2000).

In spite of advances in treatment and prevention, bacterial pathogens still pose a major threat on public health worldwide. To understand how pathogenic bacteria interact with their hosts to produce clinical disease is a fundamental issue. A key first step in this process is the identification of novel virulence determinants that may serve as targets for vaccine and drug development. In essence, the ability of pathogenic bacteria to cause disease in a susceptible host is determined by multiple virulence factors acting individually or together at different stages of infection. Virulence factors are often involved in direct interactions with the host tissues or in concealing the bacterial surface from the host's defense mechanisms (Wu *et al.,* 2008).To perpetuate the infection cycle, pathogens adhere to the host surface and gain deeper access into the host by a phenomenon termed invasion.

Invasion can be divided into two types: extracellular and intracellular. Extracellular invasion allows pathogens access to niches in tissues where they are able to proliferate, disseminate to other sites in the body, express toxins, and initiate inflammatory responses. Intracellular invasion occurs when a microbe actually penetrates the cells of a host tissue and survives within this environment. A whole lot of evidence suggests that extracellulary invading pathogens may also enter host cells and use both the extracellular and intracellular pathways during infection (Cleary & Cue 2000; Dziewanowska *et al.,* 1999; Fleiszig *et al.*, 1997). A number of Gram negative, Gram positive, and mycobacterial pathogens have the ability to enter host cells, (Finlay & Falkow 1997; Cleary & Cue 2000; Bermudez & Sangari 2000; Dehio et al., 2000) and both phagocytic and nonphagocytic cell types can serve as targets for invasion. Some pathogens have an obligate intracellular lifecycle which absolutely requires a mammalian cell for growth. These include *Chlamydia* spp, *Rickettsia* spp, and *Mycobacterium leprae* (Walker 1998). Other pathogens are facultatively intracellular, using their ability to enter and survive within host cells as a means of proliferation or spreading to other tissues.

A major advance in bacterial pathogenesis in recent years has been the identification of genes that allow pathogens to invade host non-phagocytic cells. Remarkably, these invasion genes, present in several different pathogens, were found to encode an evolutionarily related type III protein secretion pathway that serves to inject signalling proteins from the microbe into the host cell. The injected proteins then activate host cell signalling pathways that cause the host cell to internalise the microbe. These entry mechanisms are well characterised in *Salmonella spp* and *Shigella spp.* (Donnenberg 2000; Sansonetti *et al.*, 1999; Galan & Zhou 2000). A common outcome of type III secretion signalling is the rearrangement of host cell actin such that the cytoskeleton is recruited to engulf the invading microbe. Both *Salmonella* and *Shigella* engage actin regulatory proteins, called Rho GTPases, to "switch on" the actin rearrangement pathway to form nodes of actin underneath the invading pathogen (Donnenberg 2000; Galan & Zhou 2000). This type of interaction highlights the phenomenon of biochemical crosstalk between host and pathogen that is essential for penetration of host cells.

Three main forces have been found to shape genome evolution: gene gain, gene loss and gene change (Pallen & Wren 2007). Gene gain as a result of horizontal gene transfer remains the most potent source of 'innovation' and variation. However, unlike viruses, bacteria seldom acquire 'eukaryotic-like' genes from their hosts (although there seem to be some exceptions, for example, *Legionella pneumophila* (Bruggemann *et al.,* 2006). Instead, horizontal gene transfer generally occurs between different strains and species of bacteria. Bacterial genomes remain about the same size despite the pervasive effects of horizontal gene transfer, so gene gain must be balanced by gene loss (Mira *et al.,* 2001).

The most surprising snapshots of genome decay have come from recently emerged pathogens that have changed lifestyle, usually to live in a simpler host-associated niche. For example, the genomes of *M. Leprae* (Cole *et al.,* 2001), *Y. Pestis* (Parkhill 2001a) *and Salmonella enterica* serovar Typhi (Parkhill 2001b) contain hundreds or even thousands of pseudogenes; in the *M. leprae* genome, there are nearly as many pseudogenes as functional genes (Cole *et al.,* 2001).

Pathogens utilise a class of low molecular weight peptides called toxins which play an important role in pathogenesis. Toxins are analogous to biological weapons in that these are proteinaceous or non-proteinaceous molecules produced by bacteria to destroy or damage the host cell. Bacterial toxins may be broadly categorised into two distinct classes, namely endotoxin and exotoxin. Endotoxins are exemplified by LPS layer of Gram negative organisms and teichoic acid for Gram positive organisms. Exotoxins are generally enzymes which are delivered to eukaryotic cells by two different methods: (1) secretion into the surrounding milieu or (2) direct injection into the host cell cytoplasm via type III secretion systems or other mechanisms. Based on amino acid composition and function, bacterial exotoxins can be roughly categorised into the following major types- (1) A-B toxins, (2) proteolytic toxins, (3) pore forming toxins, and (4) others (Finlay & Falkow 1997). Characteristics and example of each type of bacterial exotoxins are shown below.

Type	**Characteristics**	**Example**
A-B toxins	A-B toxins have two components: the A subunit which possesses the enzymatic activity; and the B subunit which is responsible for binding and delivery of the toxin into the host cell. The enzymatic activity of the A portion of A-B toxins ranges from proteolytic activity (for example, tetanus and botulinum) to ADP ribosylating activity	cholera, pertussis, diphtheria and *P aeruginosa* exotoxin A (Merritt & Hol 1995).
proteolytic toxins	Proteolytic toxins break down specific host proteins leading to some of the characteristic clinical manifestations of the disease.	botulinum from *Clostridium botulinum*, tetanus from *Clostridium tetani (*Schiavo *et al.*, 1992), elastase (Toder *et al.*, 1991), and protease IV (Engel *et al.*, 1998) from *P. aeruginosa.*
pore forming toxins	Membrane-disrupting toxins are found in a number of bacterial species and form a pore in the host cell membrane, which ultimately leads to cell lysis. There are a growing number of pore forming toxins included in the RTX family (named for a repeat arginine (R) threonine (T) X motif within each toxin) found in many Gram negative pathogens. Although the general mechanism of pore formation and sequences are conserved in the RTX family, the target cell specificities vary. The RTX family of toxins additionally share a common method of delivery (type I secretion) (RA., 1991)	Many Gram positive bacteria contain a sulfhydryl activated cytolysin. The best characterised among these is the listeriolysin O that is necessary for the escape of *Listeria monocytogenes* from the phagosome (Andrews & Portnoy 1994).
others	These include toxins that modify host cell cytoskeleton.	immunoglobulin A (IgA) protease-type proteins (Klauser *et al.*, 1993), heat stable toxins that activate guanylate cyclise (Savarino *et al.*, 1993; Wilkins & Lyerly 1996; Aktories *et al.*, 2000; Falzant *et al.*, 1993; Oswald *et al.*, 1994).

Human are exposed to bacterial exotoxins in three ways (1) Ingestion of preformed exotoxin. The classical example is staphylocococal food poisoning and this type of bacterial disease is self-limiting, (2) Colonization of a mucosal surface folled by exotoxin production. The classical example here is the disease cholera caused by *Vibrio cholerea.* Cholera toxin stimulates hypersecretion of water and chloride ions and the patient loses massive quantities of water and gastrointestinal tract and (3) Colonization of wound or abscess followed by local exotoxin production. Example of this type is gas gangrene in which the exotoxin (α-toxin) of *Clostridium perfringens* lyes red blood cells, induces edema, and causes tissue detruction in the wound. The common theme emerging from the study of these bacterial toxins is that bacteria deploy a wide variety of strategies to disrupt host cell signalling pathways and structural integrity. This is necessary to establish and maintain infection (Wilson *et al.,* 2002).

Nearly 1000 microbial genomes have been completely sequenced till now and these include the food-borne pathogens *C. jejuni*, *S. enterica* serovar Enteritidis, *S. typhimurium*, *E. coli* O157:H7, *L. monocytogenes*, and *Shigella flexneri.* The first genome of a food-borne bacterium to be sequenced was that of *C. jejuni* and it led to the discovery of important new aspects of the biology of this organism (Parkhill *et al.,* 2000).

Plasmids play important roles in virulence and pathogenesis of disease caused by other enteric pathogens. A number of important virulence factors such as adhesions and proteases on plasmids of enterohaemorrhagic *E. coli* and enteropathogenic *E. coli* have been identified.

Enteropathogen such as *Shigella flexneri*, have evolved from a non-pathogenic *E. coli* ancestor and that the main virulence factors, such as the enterotoxin and factors required for invasion and intracellular spread were acquired by horizontal transfer of a large plasmid (Ochman & Groisman 1995; Ochman et al., 2000; Pupo et al., 2000). Five pathogenicity islands (SPI I–V) were discovered before the genome sequence became available in the food-borne pathogen *S. Typhimurium* (McClelland et al., 2001). SPI-I and SPI-II have been found in all *S. Typhimurium* strains tested and play a role in invasion and survival in macrophages, respectively. The five major PI are absent from *E. coli* suggesting that these were acquired by all serovars of *Salmonella.* It has been speculated that these SPI elements were transmitted to *Salmonella* from a common ancestor soon after its divergence from *E. coli* more than 100 million years ago (Ochman & Groisman 1996; Lee 1996). The genome sequence of *S. typhimurium* strain LT2 revealed more than sixty-two gene clusters containing four or more genes called 'islands' that are unique to certain species. Some of these 'islands' are typical PI containing putative virulence factors and sequence elements associated with their transfer from other hosts (McClelland et al., 2001) and other islands may have been acquired by horizontal gene transfer from an organism with a similar guanine plus cytosine (G+C) composition or have become adapted over a long period of time. The high number of gene clusters and islands discovered in *E. coli, Salmonella* and *Listeria* indicates that lateral gene transfer of DNA between different species and even among strains of the same species occurs more frequently than was once believed which has been a driving force in the adaptability and evolution of these pathogens. In contrast to *E. coli, Salmonella* and *Listeria,* the genome sequence of the food-borne pathogen *C. jejuni* contained very few repeat sequences and no prophages, insertion sequence elements or plasmid origins of replication despite the fact that bacteriophages and plasmids have been reported in *Campylobacter* (Lee *et al.,* 1994; Sails *et al.,* 1998). The implication is that mobile elements have played a less important role in the evolution of *C. jejuni,* perhaps because this organism is naturally competent for transformation with genomic DNA. Additionally, there is no PI or other gene clusters in the *C. jejuni* genome that has a markedly different base composition to that of the bulk of the genome. Microbial genomics initially focused on the sequencing of genomes of medically and industrially important species from diverse parts of the evolutionary tree. This approach has provided insights into the major evolutionary relationships between these microbes. As more genome sequences become available attention is turning to the comparative genomics of closely related organisms as this provides insights into the evolutionary events occurring over a shorter time scale (Ochman & Jones 2000; Ochman *et al.*, 2000). Recent genomic studies have focused on a comparison of these closely related genomes (Edwards *et al.*, 2002).

Comparative genomics has evolved as an interesting field of study. The goal of comparative genomics is to identify genetic differences across entire genomes, to correlate those differences to biological function, and to gain insight into selective evolutionary pressures and patterns of gene transfer or loss, particularly within the context of virulence in pathogenic species. Analysis of genomes from closely related species can also accelerate functional annotation of novel genes or other features (e.g., gene fusions, pseudogenes) that are apparent only in a comparative genomic context. Comparisons of the genomes can be performed either with sequence information or by using microarray-based methods to determine the presence or absence of specific genes contained on the array. However, microarray-based analysis cannot detect genes that have not yet been identified by sequencing. If a gene is present in an organism but is not on the array, then no information can be gained about that gene. The utility of cross-species genome comparisons depends on the evolutionary distance between the species. The unexpectedly high degree of intraspecies diversity in some cases suggests that a single genome sequence is not representative of the genetic inventory of a given taxonomic group but is rather a sampling of genes characterizing members of a given population in the same gene pool. Multiple strains of *Streptococcus agalactiae* have led to the concept of a pangenome, with each species consisting of a core set of genes conserved in all strains, and a dispensable genome, consisting of partially shared or unique, strain-specific genes (Raskin 2006). In recent time Bioinformatic techniques are used to analyse the genes, geneproducts etc(Kaplan and Littlejohn 2001;.

1.2. Bioinformatices technique

Codon usage analysis

For comparing genomes of pathogens, it is necessary to study their codon usage, analyze their proteomes and molecular phylogeny. Studies of codon usage can be performed using parameters like GC content, GC3 content, relative synonymous codon usage (RSCU), optimal codon anticodon energy (P2), scaled chi-square (Peden 1999),effective number of codons Nc (Wright 1990), length of the amino acid (Laa) giving the number of translatable codons (Lloyd & Sharp 1992), frequency of synonymous codons (Lsym)` (Peden 1999), codon adaptation index (CAI) (Sharp & Li 19987), frequency of optimal codons (Fop) (Ikemura 1981), codon bias index (CBI) (Chem et al., 1982), hydrophobicity, aromaticity and correspondence analysis of codon usage, RSCU and amino acid usages (Peden 1999). Codon usage study provides information of use of different codons in a genome, as it is often seen that all codons are not used evenly (Grantham *et al.,* 1981; Karlin & Mrazek 2000; Karlin *et al.,* 2001). There is dearth of comprehensive work on codon usage patterns in pathogenic food bacteria Accordingly, broad analysis of codon usage is crucial for understanding the evolution of different codon choices in these organisms. It has been postulated that major trends in codon usage patterns across genomes are determined by compositional bias, mutational pressure and/or translational selection in high or low G+C containing organisms (Knight et al., 2001). Highly expressed genes are influenced by translational selection compared to lowly expressed ones which are influenced by mutational pressure (Dos Reis, 2003). In these perspectives, studies of codon usage patterns provide a platform for better understanding of the nature of pathogenic food bacteria.

To analyze the codon usage patterns of the studied organisms the software Codon W (http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?Form=codonw) (Peden J, 1999) was used. The parameters such as GC content, GC3 content, Nc, CBI, Fop, were calculated. CAI values were calculated using the CAI calculator2 (Wu G, 2005), e-CAI server (Puigbo p, 2008). All these parameters reduce the codon usage data to a useful summary and enlighten about the factors affecting codon usage patterns in microorganisms. The GC content estimates the amount of the guanine cytosine in the nucleotide sequences. The GC3 content determines the frequency of either G or C nucleotides present in the third position of the synonymous codon. It however excludes methionine, tryptophan and the termination codons.

The effective number of codons used in a gene (Nc) is an important parameter that can measure overall codon bias of synonymous codons (Wright F, 1990). Its value represents the number of equal codons that would generate the same codon usage bias observed (Sen A, 2008).Values for the effective number of codons range from 20 (when only one codon is per amino acid) to 61 (when all codons are used in equal probability). The Nc value is influenced by mutational biases and or/selection for particular codons. However, Wright (1990) has pointed out that in organisms where mutational bias absolutely determines synonymous codon usage, the Nc value ranges between 31-61 depending upon the genomic GC content. While calculating the effective number of codons at first $\hat{F}$ (F caret) is calculated in each of the synonymous groups:

$$\hat{F} = \frac{\left(n_{aa}\sum_{i=1}^{j} p^2\right) - 1}{n_{aa} - 1} \qquad (1)$$

where, p symbolizes the fraction of usage of a codon i inside its synonymous cluster of size j, and n_{aa} the total usage of that synonymous group. The average of $\hat{F}$ for synonymous groups of same size (i.e. 2, 4, and 6) is also determined. Nevertheless, in lack of isoleucine residues $\hat{F}^3$ is calculated as an average of $\hat{F}^{av2}$ and $\hat{F}^{av4}$ and Nc value is determined using the following formula:

$$N_c = 2 + 9/\hat{F}^{av2} + 1/\hat{F}^3 + 5/\hat{F}^{av4} + 3/\hat{F}^{av6} \qquad (2a)$$

In order to include the influence of G+C bias the following equation is used to determine the expected value of Nc under random codon usage:

$$N_c = 2 + S + \{29/[S^2 + (1-S)^2]\} \qquad (2b)$$

here, S represents GC3 values.

The codon adaptation index (CAI) is a commonly used gauge to determine synonymous codon usage in prokaryotes as well as eukaryotes. It is a measure of codon usage within a gene relative to reference sets of genes that are known to be highly expressed (Sharp PM, 1987). Before calculating the codon adaptation index it is essential to determine the relative synonymous codon usage (RSCU) values from the set of highly expressed genes in an organism as determined by Sharp and Li (1987):

$$RSCU_{ij} = \frac{x_{ij}}{1/n_i \sum_{j=1}^{n_i} x_{ij}} \qquad (3)$$

here x_{ij} signifies the number of occurrence of the *j*th codon for *i*th amino acid, and n_i symbolizes the size of the synonymous group for the *i*th amino acid (i.e., 2,3,4 or 6). The CAI value for the gene is then determined by the geometric mean of the relative adaptiveness values of each of the codons present in the genes. CAI (Sharp PM, 1987) is calculated using the formula:

$$CAI = exp\left(\frac{1}{L}\sum_{k=1}^{L} \ln \omega k\right) \qquad (4)$$

where, ω_K signifies the relative adaptedness of the kth codon and L represents the number of synonymous codons in the gene. In the e-CAI server the CAI values were determined using codon usage table of the particular organism as reference. CAI values vary from 0 to 1 with higher CAI values indicating that the gene of interest has a codon usage pattern more similar to that of highly expressed genes (Sen G, 2007).

The codon bias index (CBI) (Chem, 1982) is a gauge of directional codon bias and determines the level to which a gene uses a subset of optimal codons. The codon bias index values ranges from 0 to 1. It is calculated as follows:

$$CBI=N_{opt}-N_{ran}/N_{tot}-N_{ran} \quad (5)$$

where N_{opt}= number of optimal codons; N_{tot}= total number of synonymous codons; N_{ran}=expected number of optimal codons in cases where codons are assigned randomly.

The frequency of optimal codons (Fop) (Ikemura T, 1985) is the fraction of synonymous codons that are optimal codons. If rare codons are identified there is a stipulation for determining the original Fop index (Equation 6a) or the modified Fop index (Equation 6b). All the negative values arising while determining Fop are adjusted to zero. Fop is calculated as follows:

$$F_{OP}=N_{optimal\ codons}/N_{synonymous\ codons} \quad (6a)$$

$$F_{op\ (mod)}=N_{optimal\ codons}-N_{rare\ codons}/N_{synonymous\ codons} \quad (6b)$$

where N represents the frequency of each codon type used. Fop values ranges from 0 to 1. In case where Fop values are 1 the genes are said to be made entirely of optimal codons (Ikemura T, 1985).

Laa determines the length of the amino acids. In order to test whether the values of the aforesaid indices in pathogenicity related genes, ribosomal protein genes significantly differ from that of the protein coding genes, Z test was performed.

Correspondence analysis

The Codon W software (Peden 1999) was used to calculate the correspondence analysis of codon count and amino acid usage frequencies. Owing to some fallacies associated with correspondence analysis of RSCU it was not considered for the analysis. Correspondence analysis is a multivariate statistical technique that creates a series of orthogonal axes to identify trends to explain the data variation, with each subsequent axis explaining a decreasing amount of the variation (Benzecri 1992). The file containing the gene sequences were loaded in Codon W (Peden 1999). For calculating the former the correspondence analysis menu (Menu 5) was selected. It had four options. Option 1 was used for correspondence analysis on codon count. In this option advanced correspondence analysis sub option was preferred so as to have greater control during correspondence analysis. The toggle level was changed to exhaustive; the numbers of axis altered and the program was run. Correspondence analysis on amino acid usage was performed with the help of option 3 in the correspondence analysis menu (Menu 5). Correspondence analyses on amino acid usage for the studied organisms were performed for the protein coding genes to recognize the apparent forces in characterizing adaptation of the expressed proteins.

The aforementioned parameters were correlated amongst themselves and with the principal axis of variations for correspondence analysis of codon count and amino acid usages to get further insights into their role in manipulating the diversity of codon usage patterns in the studied microorganisms.

Prediction of potentially highly expressed genes

The CAI values obtained for protein coding genes in the studied microorganisms are useful for predicting the level of expression of a gene (Sharp 1987). Wu *et al.* (2005) analyzed proteome results and validated the correlation between CAI values and expression levels showing experimentally that CAI predicted potentially highly expressed genes indeed are highly expressed. Jansen *et al.*, (2003) also confirmed this finding with yeast genome (Jansen 2003). CAI values for these studied genomes were examined to identify the predicted highly expressed genes. As defined by Wu *et al.* (Wu 2005; 2005a), the top 10% of the genes, in terms of CAI values, were classified to be predicted highly expressed genes. Clusters of orthologous groups of proteins (COG) were used to understand the functional distribution of the predicted highly expressed genes among the studied genomes. To help the analysis, each of the COG functional categories was clustered into four COG functional groups. The functional analyses of COGs based on potentially highly expressed genes in the studied organisms were used to understand their role in influencing the lifestyle of the organisms.

Phylogenetic analysis using nucleotide triplet based condensed matrix technique

The nucleotide triplet based condensed matrix phylogeny has been successfully applied in addressing the evolutionary scenario of amino-acyl tRNA synthetases in three domains of life (Mondal *et al.*, 2008) and H1N1 viruses (Sur *et al.*, 2009).

Determination of frequency of triplets of nucleic acid bases

It is well known that a DNA sequence of four letters consists of 64 possible triplets (subsequences of length 3) starting from AAA, AAT, AAG, AAC, ATA, ATT, ATG, ATC, AGA, AGT, AGG, AGC, ACA, ACT, ACG, ACC etc. The triplets contain all the relevant information for polypeptide synthesis. The introduction of a 4×4×4 cubic matrix was necessary to go for further analysis. The 4×4×4 cubic matrix comprising of 64 possible entries helps in resolving the occurrence of the probable 64 triplets in a DNA sequence. In case of a cubic matrix, it is possible to obtain three groups of 4×4 matrices each of which containing all entries of the cubic matrix. In most cases, the group $\{M_1, M_2, M_3, M_4\}$ represents the cubic matrix. The matrices were formed using all the triplets for all the studied DNA sequences. Condensed categorization of the primary sequence is derived from 4×4 matrices, whose rows and columns are related with the A, G, C and T bases. The four matrices enclose information about the frequencies of occurrence of all possible triplets of the DNA sequence along with the information regarding the frequency of occurrence of pairs of each and every letter within a DNA sequence (Randic 2001). In our method the codon positions 1, 2 and 3 were given equal weight subsequently addition or deletion of bases during the course of evolution were given due care so as to incorporate their influence. The methodology depicts DNA by condensed a matrix counting the rate of presence of adjoining base pairs (Randic 2000).

Calculation of eigen value and construction of phylogram

Leading eigenvalues were calculated using MATLAB (version 5.0.0.4069) software. These eigenvalues are a special set of scalars associated with a linear system of equations, usually matrix equations that are often regarded as characteristic roots, characteristic values (Hoffman K, 1971) (Hoffman and Kunze, 1971) and proper values or latent roots (Marcus 1988) (Marcus and Minc, 1988). Evaluation of DNA sequences for similarity or dissimilarity is normally aided by the convenience of leading eigenvectors calculated by this method. Diversity between eigenvalues was used to study sequence similarity/dissimilarity keeping in mind the characterization of a sequence by leading eigenvalue (Nandy 2006) (Nandy *et al.,* 2006). Matrices linked to each sequence are estimated and the leading eigen values computed. Variations in leading eigen values concurrent to the string are estimated and the relationships between genes investigated. Distance matrixes of the studied sequences were constructed by summing up the square of the difference of eigen values. Phylograms were built by cluster analysis of the similarity matrix using PHYLIP (Ver 3.65) (Felsenstein J, 1989) (Felsenstein 1989) and drawn with PHYLODRAW (Ver 0.8).

1.3: Molecular mechanics & force field

The first step in molecular mechanics calculation is to construct a potential energy surface (Hockney 1970) which is a function *of* atomic coordinate. The energy functions used for proteins are generally composed of bonding terms representing bond lengths, bond angles and tensional angles and non-bonding terms consisting of vander Waals interaction and electrostatic contribution. A simple molecular mechanics energy equation is given by:

$$E(R)=\frac{1}{2}\sum K_b(b-b_0)^2+\frac{1}{2}\sum K_\theta(\theta-\theta_0)^2+\frac{1}{2}\sum K_\phi\left[1+\cos(n\phi-\partial)\right]+\sum\frac{A}{r^{12}}-\frac{B}{r^6}+\frac{q_1q_2}{rD} \qquad (7)$$

The energy, E is a function of the Cartesian coordinate set, R specifying the positions of all the atoms, from which are calculated the internal coordinates for bond length (b), bond angles(θ), dihedral angles (□) and interparticle distances(r).

The first term in equation (1) represents instantaneous displacement from the ideal bond length, b_o, by a Hooke's law (harmonic) potential. Such a harmonic potential is the first approximation to the energy of a bond as a function of its length. The bond force constant k_b determines the flexibility' of the bond and can be evaluated from infrared frequencies or quantum mechanical calculations. Ideal bond length can be inferred from high resolution, low temperature crystal structures or microwave spectroscopy data. The energy associated with alteration of bond angles given by the second term in equation (7) is also represented by a harmonic potential. For rotations about bonds, torsion angle potential given by the third term in equation (7) are used. This potential is assumed to be periodic and modelled by a cosine or sum over cosine functions. The final term in equation (7) represents the contribution of non-bonded interactions and has three parts: a repulsive term preventing atoms from interpenetrating at very short distances; an attractive term accounting for the London dispersion forces between atoms; and an electrostatic term that is attractive or repulsive depending on whether the charges q_1 and q_2 are of opposite or the same sign. The first two non-bonded terms combine to give the familar Lennard-Jones 6-12 potential, which has a minimum at an interatomic separation equal to the sum of the van der Waals radii of the atoms; parameters A and B depend on the atoms involved and have been determined by a variety of methods, including non-bonding distances in crystals and gas-phase scattering measurements (Karplus, 1990)

Electrostatic interactions between pairs of atoms are represented by a Coulomb potential with D the effective dielectric function for the medium and r the distance between the two charges. Use of atomic partial charges avoids the need for a separate term to represent the hydrogen bond interaction; that is, when the positive hydrogen attached to an electronegative atom comes within van der Waals distance of a negative acceptor atom,the Coulomb attraction adds to the Lennard-Jones potential and results in a hydrogen bond.

The usefulness of empirical energy functions depends on the extent to which the parameters determined for equation (7) by the study of model systems, such as amino acids, can be employed for macromolecules, such as proteins. Evidence from a number of comparisons suggests (Blundell 1987) that this transferability condition is satisfied in many applications.

Energy minimization

For macromolecular systems, the number of local minima cost of the computations prevents exhaustive search of surface, so it is frequently impossible to determine energy minimum. There is different iterative minimization algorithm. Optimisation is the term for the mathematical process whereby the structure obtained by a series of calculation processes is compared to the starting structure and evaluated. The structure is modified to make it more consist with the parameter information within the program. Various mathematical procedures are used to determine how the geometry will change from one step to the next. For a potential energy function it is desirable to find minimum energy configuration of a system. The potential energy function of a (macro) molecular system is a very complex landscape (or *hypersurface)* in a large number of dimensions. It has one deepest point, the *global minimum and* very large numbers of *local minima,* where all derivatives of the potential energy function with respect to the coordinates are zero and all second derivatives are nonnegative. In between the local minima there are saddle points. These points are the mountain passes through which the system can migrate from one local minimum to another. All the local minima, including the global one, and of all saddle points give us the knowledge to describe the relevant structures and conformations and their free energies, as well as the dynamics of structural transitions. In particular, no minimization method exists that guarantees the determination of the global minimum in any practical amount of time. However, given a starting configuration, it is possible to find the *nearest local minimum.* Nearest in this context does not always imply nearest in a geometrical sense *(i.e.,* the least sum of Square coordinate differences), but means tile minimum that can be reached by systematically moving down the steepest local gradient. Different minimization methods are-

(A). Newton Rapson, (B). Steepest Desecent & (C). Conjugate Gradiant.

A. Newton Rapson

Many of the energy minimization programs currently in use today are based upon the mathematical principles of the Newton Rapson method. This requires first and second derivative information about the energy surface. An important property of the function used for force calculation is that they are continuous and differentiable. From simple calculus we know that the condition for a minima on a curve at point x*.the first derivative equals to zero, i.e,

$$F'(x^*) = 0$$

Now our starting point is x en we can write,

$$X^*=x+\partial x \quad (8)$$

Where ∂x represents the change which the x must undergo to reach the minimum value. The condition for the minimum can therefore be written in terms of x.

$$F'(x+\partial x) = 0$$

And expanded as a Taylor series

$F'(x +\partial x) = F'(x) + F''(x)\, \partial x + F'''(x)\, \partial x +$............

Which is also set equal to zero. Truncating the Taylor series after the set order term gives

$F'(x +\partial x) = F'(x) + F''(x)\, \partial x = 0$

Or, $\partial x = -F'(x)/ F''(x)$

Above equation can be substituted back into the equation (8)

$X^* = x -F'(x)/ F''(x)$

B. Steepest Desecent

The steepest descent method is driven purely by force gradients along potential surface. It simply takes a step in the direction of the negative gradient (hence in the direction of the force), without any consideration, of the history built up in previous steps. As the gradiant method has no information about the local curvature of the energy surface, minimization by this method slows down considerably as the gradiant decreases. Close to bottom of the potential well the energy difference can be rather small.

The principle advantage of the steepest descent method is that it is excellent at correcting major abnormalities like removal of short contacts between two non-bonded atoms at the start of the calculation. The method keeps altering the geometry until a specified cut-off value is reached and the molecule is said to be optimized. One of the drawback of this method is it is very slow to converge when the system is on a shallow potential energy surface.

C. Conjugate Gradiant

A more elegant improvement of steepest descent is that of conjugate gradients which uses information from previous steps to modify the move in the next step. In the first step, where the gradiant vector is g1 the move is given by sl =-gl

The new direction from this step takes into account the previous gradiant and follows the search direction

$S_k = -g_k + b_k s(k\text{-}l)$

where s(k -I) is the search direction from the previous step and b_k is a scaling factor. In general, steepest descents will bring you close to the nearest local minimum very quickly, while conjugate gradient brings you very close to the local minimum.

Simulated annealing is a generic probabilistic meta-algorithm for the global optimization problem. Essential feature of the simulated annealing method is that the temperature is gradually reduced as the simulation proceeds. Initially T is set to a high value and it is decreased at each step according to some annealing schedule which may be specified by the user but must end with T=O towards the end of the allotted time budget. In this way, the system is expected to wander initially towards a broad region of the search space containing solutions, ignoring small features of the energy function; then drift to low-energy regions that become narrower and narrower and finally downhill.

1.4: Molecular dynamics simulation

Molecular dynamics is the science of simulating the motions of a system of particles. It has been applied to systems as an atom and a diatomic molecule undergoing a chemical reaction, and as large as a galaxy. In all cases, the essential elements for a molecular dynamics simulation are the knowledge of the interaction potential for the particles, from which the forces can be calculated .The interaction potential, may vary from the simple gravitational interaction between stars to the complex many-body forces between atoms and molecules. Classical Newtonian equations of motion are adequate for many systems, including the biomolecules of primary concern here. But for some problems (such as reactions involving tunnelling) quantum corrections are important, and for others (such as galaxy evolution) relativistic effects may have to be included.

Two attributes of molecular dynamics simulations have played an essential part *in* their explosive development and wide range, applications. Simulations provide individual particle motions as a function of time so they can be probed far more easily than experiments to answer detailed questions about the properties of a system. Further, although the potential used in a simulation is approximate, it is completely under the user's control, so that by removing or altering specific contributions, their role in determining a given property can be examined. Computer alchemy changing the potential from that representing one system to another during a simulation -is a powerful tool for calculating free energy differences.

Simulation methods

To begin a dynamic simulation, an initial set of atomic coordinates and velocities are required. The coordinates can be obtained from X-ray crystallographic or NMR structure data, or by model-building (based on the structure of a homologous protein, for example). Given a set of coordinates, a preliminary calculation serves to equilibrate the system. The structures is first refined using an iterative minimization algorithm to relieve local stresses due to overlaps of non-bonded atoms, bond length distortions, and so on. Next, atoms are assigned velocities (v) taken at random from a Maxwellian distribution for a low temperature, and a simulation is performed for a few picoseconds. This is done by finding the acceleration a_i of atom *i,* from Newton's law $F_i = m_i\ a_i'$ (F_i' the force on the atom is computed from the derivatives of equation (7) with respect to the position; m_i is the atomic mass), and introducing it into the equation for the position r_i at time $t+\Delta t$, given r_i at time t:

$$r_i(t+\Delta t) = r_i(t)+v_i\Delta t+ \frac{1}{2}a_i(\Delta t)^2 \qquad (9)$$

The equilibration is continued by altering new velocity assignments, chosen from maxwellian distributions for temperatures that are successively increased to some chosen cho value, with intervals of dynamical relaxation. The temperature T of the system is measured by the mean kinetic energy,

$$\frac{1}{2}\sum_{i=1}^{N} m_i v_i^2 = \frac{3}{2} N k_B T$$

Where N is the number of atoms in the system, *$(v^2{}_i)$ is the* average velocity squared of the ith atom and k_B is the Boltzmann constant. The equilibration period is considered finished when the temperature is stable for longer than about 10 ps, the atomic momenta obey maxwellian distribution and different regions of the protein have the same average temperature.

Integration of the equations of motion after equilibration generates the coordinates and velocities of the atoms as a function of time. Several numerical algorithms are used to solve the equation of motion, namely: Predictor-corrector algorithm, Verlet algorithm and Leap-frog algorithm etc.

Predictor -corrector algorithm

If the classical trajectory is continuous then an estimate of the positions, velocities etc at time t+∂t may be obtained by Taylor expansion about time t:

$$r^P(t + \partial t) = r(t) + v(t)\partial t + \frac{1}{2}a(t)\partial t^2 + \ldots\ldots\ldots$$

$$v^P(t + \partial t) = v(t) + a(t)\partial t + \frac{1}{2}b(t)\partial t^2 + \ldots\ldots\ldots\ldots$$

$$a^P(t + \partial t) = a(t) + b(t)\partial t + \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (10)$$

Where r is the position, v is the velocity (the first derivative with respect to time), a is the acceleration (the second derivative with respect to time), etc. The superscript marks these as predicted values; we shall be correcting then shortly. If we truncate the expansion, retaining just the terms given in the above equation then we seem to have achieved our aim of advancing the values of the stored co-ordinates & derivatives from one time step to the next. In this example we would store four 'vectors' r,v,a,b. Equivalent alternatives would be to base the prediction, on r,v & 'old' values of the velocities $v(t-\partial t)$, $v(t-2\partial t)$. But the above equation (10) will not generate correct trajectories as time advances, as we have not introduced the equation of motion. These enter through the correction step. From the new position r^P, the forces at time $t+\partial t$ and hence the correct accelerations a^c $(t+\partial t)$. These can be compared with the predicted acceleration from equation (10)

To estimate the size of the error in the prediction step:

$$\Delta a(t+\partial t) = a^C(t+\partial t) - a^P(t+\partial t) \tag{11}$$

This error and the results of predictor step are fed into the corrector which gives;

$$\left.\begin{aligned} r^C(t+\partial t) &= r^P(t+\partial t) + c_0\Delta a(t+\partial t) \\ V^C(t+\partial t) &= v^P(t+\partial t) + c_1\Delta a(t+\partial t) \\ a^c\,(t+\partial t) &= a^P\,(t+\partial t) + c_2\Delta a(t+\partial t) \end{aligned}\right\} \tag{12}$$

The idea is that r^c $(t+\partial t)$ etc are now better approximation to the true positions, velocities etc. The general scheme of a stepwise MD simulation based on a predictor-corrector algorithm may be summarized as follow:

(a) Predict the positions velocities accelerations at time $(t+\partial t)$ using the correct values of these equation.

(b) Evaluate the forces and hence accelerations $a = f/m$, from the new position.

(c) Correct the predicted positions velocities accelerations using the new acceleration.

(d) Calculate any variables of interest such as energy, order parameter) before returning to a for the next step.

Verlet algorithm

The most widely used method of integrating the equation of motion is that initially adopted by Verlet [1967]. The method is a direct solution of the second order equations. The Verlet algorithm uses positions and accelerations at time *t* and the positions from time *t-∂t* to calculate new positions at time *t+∂t*. The Verlet algorithm uses no explicit velocities.

$r(t + \partial t) = r(t) + v(t)\partial t + \frac{1}{2}a(t)\partial t^2$

$r(t -\partial t) = r(t) - v(t) + \frac{1}{2}a(t)\partial t^2$

Summing these two equations, one obtains

$r(t + \partial t) = 2r(t) -r(t -\partial t) + a(t)\partial t^2$

The velocities are not needed to compute the trajectories, but they are useful for estimating the kinetic energy and hence the total energy. They may be obtained from the formula,

$V(t) = \{ r(t+\partial t) -r(t-\partial t)\}/2\partial t$

The advantages of the Verlet algorithm are,

i) it is straightforward and *ii)* the storage requirements are modest.

The disadvantage is that the algorithm is of moderate precision.

Leap-frog algorithm

In this algorithm, the velocities are first calculated at time *t+ 1/2dt;* these are used to calculate the positions, *r,* at time *t+dt.* In this way, the velocities *leap* over the positions, then the positions *leap* over the velocities (van Gunsteren & Berendsen 1988).

$r(t +\partial t) = r(t) +v(t +\frac{1}{2}\partial t)\partial t$

$v(t+\frac{1}{2}\partial t) =v(t-\frac{1}{2}\partial t)+a(t)\partial t$

The advantage of this algorithm is that the velocities are explicitly calculated, however, the disadvantage is that they are not calculated at the same time as the positions. The velocities at time *t* can be approximated by the relationship:

$v(t) =1/2[v\{t-\frac{1}{2}\partial t) +v(t+\frac{1}{2}\partial t)]$

Software used in molecular dynamics simulation

Different types of software are used in molecular dynamics simulation. Some common and widely used software are given in appendix I.

1.5: Homology modeling

Knowledge of a protein's tertiary structure is a prerequisite for the proper engineering of its function. Unfortunately, inspite of recent significant technological advances, the experimental determination of tertiary structure is still slow compared to the rate of accumulation of amino acid sequence data.

Without a general method for predicting the tertiary structure of an amino acid sequence, one can try to learn about the structure and function of a protein whose gene has been newly sequenced, by means of homology modeling.

Homology modeling, also known as comparative modeling of protein refers to constructing an atomic-resolution model of the "*target*" protein from its amino acid sequence and an experimental three-dimensional structure of a related homologous protein (the "*template*"). Homology modeling relies on the identification of one or more known protein structures likely to resemble the structure of the query sequence, and on the production of an alignment that maps residues in the query sequence to residues in the template sequence. It has been shown that protein structures are more conserved than protein sequences amongst homologues, but sequences falling below a 20% sequence identity can have very different structure, and reach what is called in homology modelling, the twilight zone (Chothia 1986).

The homology modeling procedure can be broken down into four sequential steps: template selection, target-template alignment, model construction, and model assessment (Marti-Renom 2000). The first two steps are often essentially performed together, as the most common methods of identifying templates rely on the production of sequence alignments; however, these alignments may not be of sufficient quality because database search techniques prioritize speed over alignment quality. These processes can be performed iteratively to improve the quality of the final model, although quality assessments those are not dependent on the true target structure are still under development.

Optimizing the speed and accuracy of these steps for use in large-scale automated structure prediction is a key component of structural genomics initiatives, partly because the resulting volume of data will be too large to process manually and partly because the goal of structural genomics requires providing models of reasonable quality to researchers who are not themselves structure prediction experts (Marti-Renom 2000).

Template selection

The preliminary task in homology modelling technique is to recognize protein structures linked to the target sequence and subsequently select those that will be used as templates (Centeno *et al.,* 2005). Position specific iterative BLAST i.e., PSI-BLAST (Altschul *et al.,* 1997) was carried out against database specification of PDB proteins which were available at the National Centre for Biotechnology Information (NCBI) Web server (http://www.ncbi.nlm.nih.gov/blast/) to find out remote similarities. The appropriate template was selected on the basis of the quality of the experimental template structure, environmental likeness and phylogenetic similarity.

Alignment of the target and template

An optimal alignment between the target sequence and template is required to construct a 3D model of the target protein, after the template sequence has been recognized. Multiple sequence alignments were performed using ClustalW 1.83 (Thompson *et al.,* 1994) using default settings and the aligned sequences were extracted in (.) PIR format (Thompson *et al.,* 1994). The aligned sequences were converted into (.) ALI format (Sali & Blundell 1993). The acquired alignments were crucially assessed in terms of number, length and position of the gaps to make it more reliable. Secondary structures were predicted using HNN (Hierarchical Neural Network method) (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_nn.html).

Construction of the rough models

The rough 3D models of the target protein from the template protein was constructed by MODELLER 9v4 program (Sali & Blundell 1993) using the alignment between target protein from the template protein. The technique is based upon the satisfaction of the spatial restraints acquired from the alignment (Centeno *et al.,* 2005). The method is theoretically comparable to that used in resolving protein structures from NMR-derived restraints. These restraints are usually acquired by assuming that the resultant distances between aligned residues in the template and the target structures are alike. These restraints obtained on the basis of homology, are generally improved by stereochemical restraints on bond lengths, bond angles, dihedral angles, and non-bonded atom-atom contacts that are attained from a molecular mechanics force field. After reducing the violation of all the restraints the model is finally obtained (Sali & Blundell 1993).

Refinement of the models

The models obtained by the homology modeling technique often contain certain amount of errors and become decisive if the concerned residues are associated with the protein function (Centeno *et al.,* 2005). To overcome this problem, refinement of the models is necessary. During the refinement process, the constructed protein models were subjected to constraint energy minimization with a harmonic constraint of 100 kJ/mol/Å^2, using the steepest descent (SD) and conjugate gradient (CG) method to remove any existing bad sectors between the protein atoms and regularizing the protein structure geometry. All of the computations were done in vacuo with the GROMOS96 43B1 parameters set using the Swiss-Pdb Viewer package (http://expasy.org/spdv/program/spdv37sp5.zip) (Kaplan and Littlejohn 2001). GROMOS is a commonly used molecular dynamics computer simulation program for investigating biomolecular systems and applied for examining conformations acquired by experimental or computer simulations (Kaplan & Littlejohn 2001). Hydrogen bonds were not considered.

Evaluation of Refined Model

For evaluation of their internal quality and reliability, the refined models were subjected to the following tests: ProSA (Wiederstein & Sippl 2007) analysis was performed to assess the accuracy and reliability of the modelled structures and check the 3D models for potential errors. VERIFY3D (Eisenberg *et al.,* 1997) was used to validate the refined structures. Here, the 3D structures of the protein models are compared to its own amino-acid sequence taking into consideration a 3D profile calculated from the atomic coordinates of the structures of correct proteins (Eisenberg *et al.,* 1997). The constructed models of the proteins were evaluated for their backbone conformation using a Ramachandran plot (Ramachandran *et al.,* 1963). The Auto Deposition Input Tool (ADIT) (http://deposit.pdb.org/validate) was used to inspect the favorable and unfavorable properties of the modeled structures. SAVES (Structure analysis and verification server) (http://nihserver.mbi.ucla.edu/SAVS/) was used to carry out the verifications of the models with PROVE and ERRAT. Presence of pockets in the structures was predicted using CASTp server (Dundas *et al.,* 2006). The refined models were submitted to ProFunc (http://www.ebi.ac.uk/thronton-srv/databases/ProFunc) (Laskowski *et al.,* 2005) to recognize the functional region in the proteins. Since there are no data on the site-directed mutagenesis, site-directed-mutagenesis predictions were carried out using the server SDM (http://www-cryst.bioc.cam.ac.uk/~sdm/sdm.php). Stability changes associated with mutations were assessed with I-Mutant 2.0 (http: //gpcr.biocomp.unibio.it/cgi/predictors/ I-Mutant2.0/I-Mutant2.0.cgi.).

Software used in Homology modeling: List of Protein Structure Prediction Software are given in Appendix II

1.6: Objectives of the research work

- Comprehensive comparative study of codon usage patterns of major food bacteria coming from different source with special reference to genes associated with pathogen. This would help in understanding the major forces influencing the codon usage patterns in these organisms.
- Analysis of codon adaptation index (CAI). It determines the resemblance between the synonymous codon usage of a gene and the synonymous codon frequency of a reference set of highly expressed genes. The CAI values will be used to predict the expression level of the genes.
- Correspondence analysis of codon usage and amino acid usage will be performed to investigate the major trends in codon and amino acid variations among the genes.
- Determination of the potentially highly expressed genes using CAI values and correlate the expression level of genes present in COG groups and study their influence on the lifestyle patterns of nitrogen fixing microbes.
- Correlation of the codon usage bias with the tRNA content of the organisms.
- Development of a novel nucleotide triplet based condensed matrix method for analyzing molecular phylogeny of some toxin genes. Characterization of the toxin genes using nucleotide triplet based condensed matrix method and construction of dendrogram using eigenvalues to determine the evolutionary relationship.
- Secondary structure prediction and determination of three-dimensional structure of important CdtB proteins using homology modeling technique so as to study the structure-function relationship.

1.7: References

Aktories K, Schmidt G & Just I (2000) Rho GTPases as targets of bacterial protein toxins. *Biol Chem.,* **381**:421–426.

Andrews NW & Portnoy DA (1994) Cytolysins from intracellular pathogens. *Trends Microbiol.,* **2**:261–263.

Bennetzen JL & Hall BD (1982) Codon selection in yeast. *J. Biol. Chem.,* **257**:3026-3031.

Benzecri JP (1992) *Correspondence analysis handbook.* [ed.] Marcel Dekker. 1992.

Bermudez L & Sangari F (2000) Mycobacterial invasion of epithelial cells. *Subcell Biochem.,* **33**:231-249.

Blundell T, Sibanda B, Sternberg M & Thornton J (1987) Knowledge-based prediction of protein structures and the design of novel molecules. *Nature,* **326**:347-352.

Bruggemann H, Cazalet C & Buchrieser C (2006) Adaptation of Legionella pneumophila to the host environment: role of protein secretion, effectors and eukaryotic-like proteins. Curr Opin Microbio., **9**:86–94.

Casadevall A & Pirofski LA (1999) Host-Pathogen Interactions: Redefining the Basic Concepts of Virulènce and Pathogenicity. *American Society for Microbiology,* 67:3703-3713.

Chothia AM & Lesk C (1986) The relation between the divergence of sequence and structure in proteins. *EMBO. J.,* **5**: 823-826.

Chugh TD (2008) Emerging and re-emerging bacterial diseases in India. *J. Biosci.* **33**:549-555.

Cleary P & Cue D (2000) High frequency invasion of mammalian cells by β hemolytic streptococci. *Subcell Biochem.,* 33:137–66.

Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D ***et al.,*** (2001) Massive gene decay in the leprosy bacillus. *Nature.* **409**:1007–1011.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE *et al.*, (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature,* **393:**537-544.

Dehio C, Gray-Owen S & Meyer T (2000) Host cell invasion by pathogenic Neisseriae. *Subcell Biochem.,* **33**:61–96.

Domenighini M, Magagnoli C, Pizza M & Rappuoli R (1994) Common features of the NAD-binding and catalytic site of ADP-ribosylating toxins. *Mol Microbiol.***14**:41–50

Donnenberg MS (2000) Pathogenic strategies of enteric bacteria. *Nature,* 2000 **406**:768–74.

Dos Reis M, Wernisch L & Savva R (2003) Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Res.,* **31**:6976-6985.

Dziewanowska K, Patti JM, Deobald CF, Bayles KW, Trumble WR & Bohach GA, (1999) Fibronectin binding protein and host cell tyrosine kinase are required for internalization of Staphylococcus aureus by epithelial cells. *Infect Immun.* **67:**4673-4678.

Edwards AE, Olsen GJ & Maloy SR (2002). Comparative genomics of closely related salmonellae. *Trends in Microbiology.* **10**:94–99.

Engel LS, Hill JM, Caballero AR, Green LC & O'Callaghan RJ (1998) Protease IV, a unique extracellular protease and virulence factor from Pseudomonas aeruginosa. *J Biol Chem.,***273**:16792–16797.

Falzano L, Fiorentini C, Boquet P, Donelli G (1993) Interaction of Escherichia coli cytotoxic necrotizing factor type 1 (CNF1) with cultured cells. *Cytotechnology.* **11**:S56–8.

Felsenstein J (1989) PHYLIP-Phylogeny Inference Package (version 3.2). *Cladistics,***5**:164-166.

Finlay BB & Falkow S (1997) Common themes in microbial pathogenicity revisited. *Micro Mol Biol Rev.,* **61**:136-169.

Fleiszig SM, Wiener-Kronish JP, Miyazaki H, Vallas V, Mostov KE, Kanada D, Sawa T, Yen TS & Frank DW (1997) *Pseudomonas aeruginosa*-mediated cytotoxicity and invasion correlate with distinct genotypes at the loci encoding exoenzyme S. *Infect Immun* ., **65**:579–586.

Fujii G, Choe S, Bennett MJ & Eisenberg (1991) Crystallization of diphtheria toxin. *J.Mol Biol.* **222**:861–864.

Galan JE & Zhou D (2000) Striking a balance: modulation of the actin cytoskeleton by salmonella. *Proc Natl Acad Sci U S A.* **97**:8754–61.

Garcia E, Llull D & Lopez R (1999) Functional organization of the gene cluster involved in the synthesis of the pneumococcal capsule. *Int Microbiol.* **2**:169–176.

Grantham R, Gautier C, Gouy M, Jacobzone M & Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* 9:r43-r74.

Hacker J & Kaper J (2000) Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol.* 54:641–79.

Hall G, Kirk MD, Becker N, Gregory JE, Unicomb L, Millard G, Stafford R, Lalor K, and the OzFoodNet Working Group (2005). Estimating Foodborne Gastroenteritis, Australia. *Emerging Infectious Diseases.* 8:1257-1264.

Hall GV, D'Souza RM & Kirk MD (2002). Foodborne disease in the new millennium:out of the frying pan and into the fire? *Med J Aust.,***177**:614–8.

Hamood A, Wick M, Iglewski B (1990) Secretion of toxin A from Pseudomonas aeruginosa PAO1, PAK, and PA103 by Escherichia coli. *Infect Immun.* **58**:1133–40.

Hockney RW (1970) *Methods in Computational Physic.* Alder B, Fernbach S & Rotenberg, M. (eds.) Vols. 9, Academic Press Plasma Physics New York

Hoeprich PD (1983) Host-parasite relationships and the pathogenesis of infectious disease. *Infectious diseases.* In P. D. Hoeprich. (ed.).Harper & Row

Hoffman K, Kunze R (1971) Linear algebra, Englewood C. 2nd eds. Prentice-Hall, New Jersey

Ikemura T (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons of its protein genes. *J Mol Bio.***146**:1-21.

Ikemura T (1985). Codon usage and transfer-RNA content in unicellular and multicellularorganisms. *Mol Biol Evol.* **2**:13-34.

Jansen R, Bussemaker HJ & Gerstein M (2003) Revisting the codon adaptation index from a whole-genome perspective: analyzing the relatioship between gene expression and codon occurrence in yeast using a varity of models. *Nucleic Acids Res.*31:2242-2251.

Kaczanowski S & Zielenkiewicz P (2010) Why similar protein sequences encode similar three-dimensional structures? *Theoretical Chemistry Accounts.***125**:543-50.

Karlin S & Mrazek J (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol.* **182**:5238-5250.

Karlin S, Mrazek J, Campbell A & Kaiser D (2001) Characterizations of highly expressed genes of four fast-growing bacteria. *J Bacteriol.* **183**:5025-5040.

Karplus M & Petsko G (1990) Molecular dynamics simulations in biology. *Nature.* **347**:631–639.

Klauser T, Pohlner J & Meyer TF (1993) The secretion pathway of IgA protease-type proteins in gram-negative bacteria. *Bioessays.* **15**:799–805.

Klose KE (2001) Regulation of virulence in Vibrio cholerae. *Int J Med Microbiol.* **291**:81–88.

Knight RD, Freeland SJ & Landweber LF (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol 2: research 00101-001013;.* 2001, Vols. research 00101-001013;.

Lee CA (1996) Pathogenicity islands and the evolution of bacterial pathogens. *Infectious Agents and Disease.* **5**:1-7.

Lee CY, Tai CL, Lin SC & Chen YT (1994) Occurrence of plasmids and tetracycline resistance among Campylobacter jejuni and Campylobacter coli isolated from whole market chickens and clinical samples. *International Journal of Food Microbiology.* **24**: 161–170.

Lindsay J, Ruzin A, Ross HF Kurepina N & Novick RP (1998) The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in Staphylococcus aureus. *Mol Microbiol.* **29**:527–43.

Lloyd PM & Sharp AT (1992) CODONS - a microcomputer program for codon usage analysis. *J Heredity.* **83**:239-240.

Marcus M & Mine H (1988). *Introduction to Linear Algebra.* Dover

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F & Sali A (2000) Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct.* **29**:291-325.

McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P,Courtney L, Porwollik S, Ali J, Dante M, Du F, Hou S & Layman D *et al.,* (2001) Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. *Nature.* **413**:852–856.

Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C,Griffin PM & Tauxe RV(1999) Food-related illness and death in the United States. *Emerging Infectious Diseases.* **5**:607-625.

Merritt EA & Hol WG (1995) AB5 toxins. *Curr Opin Struct Biol.* **5**:165–71.

Mira A Ochman H & Moran NA (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**:589–596.

Nakao H & Takeda T (2000). Escherichia coli Shiga toxin. *J Nat Toxins.* **9**:299-313.

Nandy A, Harle M & Basak SC (2006) Mathematical descriptors of DNAsequences: Development and applications. *ARKIVOC.* **IX**:211-238.

Ochman EA & Groisman H (1996) Pathogenicity islands: bacterial evolution in quantum leaps. *Cell.* **87**:791–794.

Ochman EA & Groisman H (1995). The evolution of invasion by enteric bacteria. *Canadian Journal of Microbiology.* **41**:555–561.

Ochman H & Jones IB (2000) Evolutionary dynamics of full genome content in Escherichia coli. *EMBO Journal.* **19**:6637–6643.

Ochman H & Moran N (2001) Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science.* 292:1096–1099.

Ochman H, Lawrence JG & Groisman EA (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature.* 2000, Vol. 405, pp. 299–304.

Oswald E, Sugai M, Labigne A, Wu HC, Fiorentini C boquet P & O'Brien AD (1994) Cytotoxic necrotizing factor type 2 produced by virulent Escherichia coli modifies the small GTP-binding proteins Rho involved in assembly of actin stress fibers. *Proc Natl Acad Sci U S A.* **91**:3814–3818.

Paden J (1999) Analysis of codon usage. *PhD thesis*, University of Nottingham

Pallen MJ & Wren BW (2007) Bacterial pathogenomics. *NATURE.* **449**:835-842.

Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MTG, Sebaihia M, Baker S, Basham D, Brooks K, Chillingworth T, Connerton P, Cronin A, Davis P, Davies RM, Dowd L, White N, Farrar J, Feltwell T, Hamlin N, Haque A, Hien TT, Holroyd S, Jagels K, Krogh A, Larsen TS, Leather S, Moule S, Ó'Gaora P, Parry C, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S & Barrell BG (2001) Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. *Nature.* **413**:848–852.

Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T & Holroyd S et al (2000) The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. *Nature,* **403**:665–668.

Parkhill J, Wren BW, Thomson NR, Titball WR, Holden MTG, Prentice MB, Sebaihia M, James KD, Churcher C & Mungall KL et al., (2001) Genome sequence of Yersinia pestis, the causative agent of plague. *Nature.* 2001, Vol. **413**:523–527.

Puigbo P, Bravo IG , vallve-Garcia S (2008) E-CAI: a novel server to estimate an expected value of codon adaptation index(e CAI). *BMC Bioinformatics.* **9**:65 doi:10.1186/1471-2105-9-65

Pupo GM, Lan R & Reeves PR. (2000). Multiple independent origins of Shigella clones of Escherichia coli and convergent evolution of many of their characteristics. *Proceedings of the National Academy of Sciences USA.* **97**:10567–10572.

Randic M (2000) Condensed representation of DNA Primary sequences. *J. Chem Inf Comput Sci.* **40**:50-56.

Randic M, Guo X, Basak SC (2001) On the Characterization of DNA Primary sequences by triplet of Nucleic Acid Bases. *J. chem.Inf. Comput Sci.* **41**:619-626.

Raskin DM, Seshadri R, Pukatzki SU, Mekalanos JJ (2006). Bacterial Genomics and Pathogen Evolution. *Cell.* DOI 10.1016/j.cell.2006.02.002.

Sails AD, Wareing DR, Bolton FJ, Fox AJ & Curry A (1998) Characterisation of 16 Campylobacter jejuni and C. coli typing bacteriophages. *Journal of Medical Microbiology.* **47**:123–128.

Sansonetti P, Tran Van Nhieu G, Egile C (1999). Rupture of the intestinal epithelial barrier and mucosal invasion by Shigella flexneri. *Clin Infect Dis.* **28**:466–75.

Savarino SJ, Fasano A, Watson J, Martin BM, Levine MM, Guandalini S & Guerry P (1993) Enteroaggregative Escherichia coli heat-stable enterotoxin 1 represents another subfamily of E coli heat-stable toxin. *Proc Natl Acad Sci U S A.* 90:3093–3097.

Schiavo G, Benfenati F, Poulain B, Rossetto O, Polverino P, Laureto D, dasgupta BR & Montecucco C (1992) Tetanus and botulinum-B neurotoxins block neurotransmitter release by proteolytic cleavage of synaptobrevin. *Nature.***359**:832–835.

Sen A, Sur S, Bothra AK, Benson DR, Normand P, Tisa LS (2008). The implication of life style codon usage patterns and predicted highly expressed genes for three Frankia genomes. *Anton van Leewen.* 93:335-346.

Sen G, Sur S Bose D, Mondal U, Furnholm T Bothra AK, Tisa LS Sen A (2007) Analysis of codon usage pattern and predicted highly expressed genes for six phytopathogenic xanthomonas genomes shows a high degree of conservation. *InSilico Bio.* **7**:547-558.

Sharp PM & Li WH (1987) The codon adaptation index: A measure of directional synonymous codon usage bias, and its potential applications. *Nucl Acids Res* 15:1281-1295.

Smith H (1977) Microbial surfaces in relation to pathogenicity. *Bacteriol. Rev.* **41**:475-500.

Stein PE, Boodhoo A, Armstrong GD, Cockle, SA, Klein MH & Read RJ (1994) Structure of a pertussis toxin-sugar complex as a model for receptor binding. *Nat Struct Biol.***1**:591–596.

Sudershan RV, Rao P & Polasa K (2009) Food safety research in India: a review. *As. J. Food Ag-Ind.* **2**:412-433.

Swaminathan B & Feng P (1994). Rapid detection of food-borne pathogenic bacteria. *Annu. Rev. Microbiol.* 48:401-426.

Toder DS, Gambello MJ, Iglewski BH (1991) Pseudomonas aeruginosa LasA: a second elastase under the transcriptional control of lasR. *Mol Microbiol.* **5**:10.

Walker T (1998) *Microbiology.* Philadelphia: WB Saunders Company

Watson DW & Brandly CA (1949) Virulence and pathogenicity. In C. E. Clifton, S. Raffel & H. A. Barker. (ed.).Annual reviewof microbiology.Annual Reviews, Inc.

Welch RA (1991) Pore-forming cytolysins of gram-negative bacteria. *Mol Microbiol.* **5**:521–528.

Wilkins TD & Lyerly DM (1996) Clostridium difficile toxins attack Rho. *Trends Microbiol.* **4**:49–51

Williamson AR (2000) Creating a structural genomics consortium. *Nat Struct Biol.*7: 953

Wilson JW, Schurr MJ, LeBlanc CL, Ramamurthy R, Buchanan KL, Nickerson CA. (2002) Mechanisms of bacterial pathogenicity. *Postgrad Med J.***78**:216–224.

Wren B (2000) Microbial genome analysis: insights into virulence, host adaptation and evolution. *Nat Rev Genet.*1:30-39.

Wright F (1990). The effective number of codons used in gene. *Gene.* **87**:23-29.

Wu G, Culley DE & Zhang W (2005) Predicted highly expressed genes in the genomes of Streptomyces coelicolor and Streptomyces avermitilis and the implications for their metabolism. *Microbiology.* 151: 2175-2187.

Wu HJ, Wang AHJ & Jennings MP (2008) Discovery of virulence factors of pathogenic bacteria. *Current Opinion in Chemical Biology.* **12**:93-108.

Wu G, Nie L & Zhang WW (2005a). Predicted highly expressed genes in Nocardia forcinica and the implication for its primary metabolism nocardial viulence. *Anton Van Leeuwen.* **89**:135-146.

Youmans GP, Paterson PY & Sommers HM (1975) *The biologic and clinical basis of infectious disease.* W. B. Saunders Co.

Ziebuhr W, Ohlsen K, Karch H, Korhonen T& Hacker J (1999) Evolution of bacterial pathogenesis. *Cell Mol Life Sci.* **56**:719-28

Chapter II

Review of Literature

2.1: Discovery of pathogens and pathogenicity Island

Cholera is a sometimes fatal disease caused by infection with the bacteria *Vibrio cholerae* (Ryan 2004) which is transmitted through contaminated water. Prior to the discovery of an infectious cause, the symptoms of cholera were thought to be caused by an excess of bile in the patient. The disease Cholera gets its name from the Greek word *choler* meaning bile. This was consistent with medical thought at the time, which held that four liquids or *humors* controlled health, and lead to such medical practices as bloodletting as a method of curing illnesses. The bacterium was first reported in 1849 by M. Gabriel Pouchet, who discovered it in stools from patients with Cholera. But he did not appreciate its significance (William 1979). The first scientist to understand the significance of *Vibrio cholerae* was an Italian anatomist Filippo Pacini, who published detailed drawings of the organism in "Microscopical observations and pathological deductions on cholera" in 1854. He would go onto publish additional papers in 1866, 1871, 1876 and 1880, all of which were ignored by the scientific community. He correctly described how the bacteria caused diarrhea, and developed treatments that were found to be effective (Bentivoglio & Pacini 1995). But his findings did not influence medical opinion. In 1874, scientific representatives from 21 countries voted unanimously to resolve that Cholera was caused by environmental toxins from *miasmatas,* or clouds of unhealthy substances which float in the air (Howard-Jones N 1984).

Reports of pathogenic E. coli appear in medical literature as early as 1947 (Ruchman & Dodd 1947) Publications regarding variants of E. coli which cause disease appeared regularly in medical journals throughout the 1950s, 60s, and 70s, (Macqueen 1954; Mc-Clure 1955; Gronroos 1957; Cowart & Thomason 1965; Linde *et al.,* 1966) with fatalities being reported in humans and infants starting in the 1970s (Glantz 1970; Drucker *et al.,* 1970; Smith & Gyles 1970). Infection with the bacteria *Helicobacter pyoli* is the cause of most stomach ulcers. The discovery is generally credited to Australian gastroenterologists Dr. Barry Marshall and Dr. J Robin Warren, who published their findings in 1983. The pair received the Nobel Prize in 2005 for their work. Before this, nobody really knew what caused stomach ulcers, though a popular belief was that the "stress" played a role. Some researchers suggested that ulcers were a psychosomatic illness (Paulley 1975; Kellner 1975; Aitken & Cay 1975). Besides, a number of food borne pathogens causing bacteria like *Salmonella* (non-typhoidal), *Listeria monocytogenes*, *Campylobacter*, (Mead *et al.,* 1999) were discovered. Genomic islands may contain large blocks of virulence determinants (adhesins, invasins, toxins, protein secretion systems, antibiotic resistance mechanisms, etc) and thus are referred to as pathogenicity islands. Pathogenicity islands were first described in pathogenic species of *E coli*, but have since been found in the genomes of numerous bacterial pathogens of humans, animals, and plants (*Salmonella*, *Vibrio*, *Shigella*, *Yersinia*, *Listeria*, *S aureus*, etc) (Garcia et al., 1999; Lindsay *et al.,* 1998). Pathogenicity islands are believed to have been acquired as a block by horizontal gene transfer because of their G+C content is significantly different from that of the genomes of the host micro-organism and they are often flanked by direct repeats. Many scientists are currently beginning to acquire understanding of the molecular mechanisms involved in the action of toxins. The good news is that there are a limited number of major toxin families that display common structural and biochemical motifs which may be exploited for future therapeutic development and these may be effective against multiple organisms.

2.2: Genetics of food pathogen

In 1906 the term "genetics" is introduced. Various techniques like mutations, deletion mapping, cloning vectors etc. have facilitated the identification of genes associated with food pathogen. The first genome of a food-borne bacterium to be sequenced was that of *C.*

jejuni and it led to the discovery of important new aspects of the biology of this organism (Parkhill *et al.,* 2000). PAIs were first discovered in pathogenic *E. coli.* Pathogenicity islands of various pathogens are cag (*H. pylori),* SPI1-5 (*Salmonella* spp) VPI (*V. cholerae) etc* are responsible for different diseases (Hentschel & Hacker 2001). Kim *et al.,* design a method for the Microarray detection of food-borne pathogens using specific probes prepared by comparative genomics (Kim *et al.,* 2008)

2.3: Research trends in food pathogen bacteria (FPB)

The discovery of pathogenic organisms throughout the globe, studies concerning their gene products as well as the development of molecular biology helped FPB research enter a new stage where basic research combined with latest techniques. Techniques such as immune magnetic separation (IMS) and polymerase chain reaction (PCR) have paved the way for rapid and sensitive detection of foodborne pathogens, and advances in nanobiotechnology have allowed for miniaturization of devices. Collaborations between workers in the fields of engineering, nanotechnology and food science have introduced new lab-on-a-chip technologies permitting development of portable, hand-held biosensors for food pathogen detection. Despite the recent advances in food pathogen detection, there still exist many challenges and opportunities to improve the current technology.

2.4: Beginning of interdisciplinary research and dawn of bioinformatics

It was in the year 1986 that the Department of Energy (DOE), USA and National Institutes of Health (NIH) started the Human Genome Project (HGP) and became one of the most happening experiments in the late 20th century. The objective of the project was to identify all of the genes in humans and craft a database containing the information (Ideker *et al.,* 2001). A number of other genome projects also started in major industrialized countries of Europe and Japan. The scientists were at first skeptical about the HGP, since huge amount of money had to be spent that would hamper basic research and as the project moved scientists worried about the massive amount of data and its interpretation (Bloom, 2001) In the year 2000, Prof. Collins of the National Human Genome Research Institute and Prof. Craig Venter of Celera Genomics appeared in a press conference and stated that they had achieved what was thought impossible and published the draft sequence of human genome (Wade 2000). That was the beginning and it opened the flood-gates for other genome sequencing projects. Gradually sequences of mouse, rat, worms and yeast were completed (Miller *et al.,* 2004). Kyrpides (1999) reported that in the end of the 20th century, there were 24 complete genomes that included 16 bacterial, 6 archaeal, and 2 eukaryotic genomes and currently there are more than two thousand genomes available in public databases (Kyrpides 1999). The large numbers of genomes resulted in the generation of huge amount of information concerning the genetic nature of biological organisms spanning different kingdoms, groups and lineages etc. Bloom proposed that the greatest problem appeared to be the interpretation of underlying information from genomes leading to materialization of the new science of bioinformatics. Bioinformatics revolutionized the science of biology and directed it towards a more holistic approach compared to the reductionism visible in molecular biology research in the late 20th century (Bloom 2001). Now an organism is viewed as a system comprising of the information associated with genes and proteins that are responsible for maintaining day to day functions and networks of regulations that spell out how gene expression occur (Ideker *et al.,* 2001). In 1995, the publication of the complete genome of *Haemophilus influenzae* marked the beginning of another revolution in the field of bioinformatics (Fleischmann *et al.,* 1995). Currently there are about two thousand bacterial and archaeal genomes in the public domain. The publication of huge amount of sequence data helped in the development of high end computers, smart computing tools, for large-scale annotation, functional classification of the proteins (Searls 2000) and development of specific databases (Birney *et al.,* 2002) for availability to the broad scientific community. As the science of bioinformatics developed, computation became cheaper and cheaper and was duly complemented with the growth of Internet since the late 90s (Perez-Iratxeta *et al.*, 2006). High throughput tools greatly developed in the beginning of the 21st century as genetic data became a gold mine for researchers (Perez-Iratxeta *et al.*, 2006). New software started developing for more efficient and comprehensive analysis of the genomes, proteomes and proteins. The stage was set and bioinformatics had become the leading science of the 21st century.

2.5: Post genomic era and FPB

The amalgamation of the knowledge of physiology, biochemistry, genetics and molecular biology gave idea about the understanding of the mechanism of host-pathogen interaction in pre-genomic era. The completion of the genomes of *Salmonella enterica* (McClelland *et al.*, 2001), *Helicobacter pylori Strain* G27 (Baltrus *et al.*, 2009) and sequences for a number of pathogenic bacteria. Pathogenicity islands were first described in pathogenic species of *E coli*, but have since been found in the genomes of numerous bacterial pathogens of humans, animals, and plants (*Salmonella*, *Vibrio*, *Shigella*, *Yersinia*, *Listeria*, *S aureus*, etc) (Garcia *et al.*, 1999; Lindsay *et al.*, 1998). The first genome of a food-borne bacterium to be sequenced was that of *C. jejuni* and it led to the discovery of important new aspects of the biology (Parkhill *et al.*, 2000). The studies on the genomes exposed new evidences pertaining to evolution and structure, interactions between host-pathogen interactions.

The research on functional genomics and proteomics for the food pathogenic bacteria is of great significance in the 21st century. It has become necessary to focus much on the comparative codon usage patterns, whole proteome analysis and molecular phylogeny using bioinformatics tools. Comparative genomics particularly focusing codon usage using different parameters is expected to provide insight into the inherent molecular nature of the genomes of food pathogenic bacteria.

2.6: Previous works on codon usage patterns

In the post genomic era increasing number of genomes put forth a concept among computational biologists that each and every genome has its own story. Since the time when the first nucleic acid sequences were obtained a number of hypothesis on the evolution of genomes have been put forward. The genetic code has been one of the most interesting aspects of biological science. The code is degenerate with multiple codons coding for a particular amino acid. Groups of codons coding for a particular amino acid are synonymous ones. It has been reported that these synonymous codons are somewhat conserved across species (Peden 1999). The increase in sequence information *albeit* partial in the 1980's facilitated the studies concerning the usage of synonymous codons of organisms. Majority of work on codon usage patterns at that period focused upon *E.coli* (Peden 1999). Gradually the techniques for codon usage were applied upon mammalian, bacterial, bacteriophage, viral and mitochondrial genes (Grantham *et al.*, 1980a; Grantham *et al.*, 1981; Grantham *et al.*, 1980b). On the basis of studies conducted on mRNAs from a number of prokaryotic and eukaryotic species, Grantham *et al.* (1980a) proposed the "Genome Hypothesis" which hypothesized that codon usage pattern of a particular genome was an explicit attribute of that organism. Grantham *et al.*, (1981) reported that difference in codon usage pattern might be associated with the tRNA content. More work on the codon usage patterns (Gouy & Gautier, 1982) in *E.coli* regarding codon usage and tRNA abundance led to the conclusion that highly expressed genes exhibited non-random codon usage and used a small set of codons that corresponded to abundant tRNAs. However, it was not clear why specific synonymous codons were used preferably. Grosjean & Fiers (1982) opined that optimal codon choice is the outcome of the necessity imposed by interaction between codon and cognate tRNA. Ikemura (1981) defined optimal codon as one that was translated by the most abundant cognate tRNA which he later amended (Ikemura, 1985). These optimal codons are under the influence of translational efficiency. Kurland (1991) reported that translational efficiency is shaped by highest turnover of ribosomes, effectiveness of aminoacyl-tRNA harmonizing and ternary complex conditions. Sharp *et al.* (1993) reported that the preference of some synonymous codons were the outcome of translational selection i.e., for increasing efficiency and accuracy a codon is used that is translated by the abundant tRNA species. Rocha (2004) correlated codon usage bias from the tRNA point of view. He proposed that co-evolution of tRNA gene composition and codon bias in genomes from tRNA's point of view concur with the selection-mutation-drift theory. A number of studies (Sharp *et al.*, 1993; Carbone *et al.*, 2005) revealed that codon bias is influenced by effective population size, translational selection, mutational pressure, compositional bias and genetic drift. Chen *et al.*, (2004) postulated that codon bias is first and foremost influenced by mutational pressure and then translational selection. Studies on translational selection helped in the detection of highly expressed genes in genomes (Karlin & Mrazek, 2000). Majority of the highly expressed genes are associated with cell growth and cell division. It has been found that evaluating translational selection in quite difficult in genomes with a high or low G+C content owing to the consequence of intense levels of G+C content on codon usage. Medigue *et al.*, (1991) applied the principles of codon usage, cluster analysis and correspondence analysis to study horizontal gene transfer mechanisms. A number of workers have suggested that codon usage and amino acid usage (Goldman & Yang 1994; Nesti *et al.* 1995; Pouwels & Leunissen 1994; Schmidt 1995) helps in determining the phylogenetic relationships between organisms. Besides, codon usage amino acid usage has also been studied at large (Peden 1999). Ikemura (1981) was instrumental in reporting good correlation between amino acid composition and codon bias. Hydrophobicity, aromaticity and amino acid charges are testified to be influencing amino acid usage (Lobry & Gautier, 1994). The prediction of open reading frames has been performed utilizing the information of codon usage (Krogh *et al.* 1994; Borodovsky *et al.* 1995). Peden (1999) provided some classic examples of the utilization of GeneMark prediction programme (Borodovsky *et al.*, 1994) for identification of coding sequences from shotgun genome sequencing projects.

Sharp and Li (1987) were the pioneers in developing the Codon Adaptation Index (CAI) to assess the similarity amid the synonymous codon usage of a gene to that of the reference set. This particular index has been commonly used as a parameter for predicting the highly expressed genes in an organism. Besides CAI, a number of indices are regularly used to investigate codon usage patterns in a number of organisms. Peden (1999) opined that two types of indices one evaluating on the whole deviation of codon usage from the expected set and the other determining bias towards a specific subset of optimal codons are used. Indices such as GC content, GC3 content (Peden 1999) effective number of codons (Nc) (Wright 1990), relative synonymous codon usage (RSCU) (Sharp *et al.,* 1986) scaled chi-square and G statistic, P2 index (Gouy & Gautier, 1982) measuring the codon-anticodon interaction, codon bias index (CBI) (Bennetzen & Hall, 1982), Fop (frequency of optimal codons) (Ikemura, 1985) and codon adaptation index are very significant in studies concerning codon usage patterns. Besides these indices, correspondence analysis (Benzecri, 1992) a type of multivariate statistical analysis are commonly used to find out the degree of associations between different genes and amino acids. It is a technique that explores non random synonymous codon usage.

This important field of study starved for good software in the 1980's. The basic problem was the integration of different indices into one program that would serve the researchers well. The foremost software was CODONS (Lloyd & Sharp, 1992). The GCG package (GCG, 1994) consisted of programs for analysis codon usage. CORRESPOND was used for performing multivariate statistical analysis (Peden 1999). NetMul (Thioulouse *et al.,* 1995) was designed as a subset of ADE software for studying of codon usage and multivariate statistical analysis. CODON W (Peden 1999) took over the stage and became very popular because of its robust nature and error free analysis. INCA (Supek & Vlahovicek, 2004) is also being used by researchers. Software was also developed for codon usage optimization. This involved alteration of rare codons in target genes so as to imitate the codon usage of host with no modification of the amino acid sequence of encoded proteins (Gustafsson *et al.*, 2004). GeneDesign (Richardson *et al.*, 2006), Synthetic Gene Designer (Wu *et al.*, 2006), Gene Designer (Villalobos *et al.*, 2006) are some of the software providing a platform for synthetic gene design and codon optimization. For estimation of the highly expressed genes using CAI there has been a number of software. CODON W (Peden 1999), JCAT (Grote *et al.*, 2005), CAI CALCULATOR (Wu *et al.*, 2005), ACUA (Umashankar *et al.*, 2007) and e CAI server (Puigbo *et al.*, 2008) are commonly used. However, the e CAI server (Puigbo *et al.*, 2008) has proved to be the most powerful and efficient tool for estimation of expression levels of the genes.

Although a lot of work has been performed and is going on in codon usage of different microorganisms but very little work performed on codon usage of food pathogenic bacteria. The first species in which codon usage was examined in detail, the bacterium Escherichia coli (Post & Nomura 1980; Ikemura 1981) and the yeast Saccharomyces cerevisiae (Ikemura 1982; Bennetzen & Hall 1982), were both found to show strong evidence of natural selection on codon usage. All these are just very little and a comprehensive work with focus on lifestyle patterns are the need of the hour.

2.7: Phylogenetic studies on FPB

Nothing in biology makes sense without evolution. This also applies for food pathogenic bacteria. To establish the differentially evolved genes Eswarappa *et al.*, compared the phylogeny of the nine potential differentially evolved genes with phylogeny of the S.

eneterica species and with the phylogeny of the five pathogenecity island of Salmosella (SPI-1 to SPI-5)(Eswarappa *et al.,* 2008) . This also applies for other pathogenic bacteria.

2.8: Structural bioinformatics of food pathogenic bacteria

High throughput sequencing of a number of eukaryotes and a number of bacteria has developed the science of structural bioinformatics. Structural bioinformatics is expected to give rise to a large number of protein structures. The technological benefits of this field are already yielding results and have a substantial impact on structural biology research for prokaryotes (Burley & Bonnano, 2002).Metropolis et al., introduced the Monte-Carlo technique to the solving of physical equations. It described the idea of using random numbers to project a representative subset of conformational space, whilst using the exponent of the energy as a probability filter (Metropolis *et al.,* 1953). Levitt and Warshel simulated the folding of the Bovine Pancreatic Trypsin Inhibtor (Levitt & Warshel 1975). Case and Karplus work on “Dynamics of ligand binding to heme protein” in 1979. This is arguably the first simulation of ligand moving through the protein. At this early stage in the game, they could either fix the protein and watch the oxygen bounce around, or let individual sidechains get hit by the oxygen. Each oxygen atom were simulated for 3.75 ps. For these pioneers it was a surprise to see that the oxygen bounces around the inside of the myoglobin, without getting too far. Nevetheless, they identified 2 different pathways for the oxygen to travel into the binding site(Case & Karplus 1979). Umbrella sampling is the most popular method of exploring large conformational changes in MD. In 19982Notthru p et al show “Dynamical theory of activated processes in globular proteins”. In this paper, Karplus and friends modeled a rather more modest conformational change: the swinging of an aromatic residue sidechain. From the simulations, they generated a free-energy surface, from which they calculated a sidechain flipping rate. This paper is important not just for simulating the first sidechain flip but also for introducing proteins to "umbrella sampling"(Northrup *et al.,*1982). First application of normal modes to identify low-frequency oscillations using the energy minimum of the molecular mechanics force-field of a protein deseribe by (Brooks & Karplus 1983). This is the basic technique to identify domain-level motions in a protein. First simulation of a protein in explicit waters was done by Levitt & Sharon (Levitt & Sharon 1988). Suddenly, acceptable computer resources got a whole lot more expensive. Guilbert et al gave first description of the RMSD potential, a powerful method for identifying low-energy pathways in the neighborhood of a given static structure (Guilbert *et al.,* 1995). Israelachvili & Wennerström shows that water molecules can have structuring effects of several Ångstroms. Flags the importance of using explicit water molecules (Israelachvili & Wennerström 1996). The first reported 1 microsecond MD simulation, it was a mamoth effort for the late 90's, really pushing the technology of parallel clusters, a technology which we all pretty much take for granted now. They tried to fold a tiny protein, the villin headpiece subdoamin, and got some of the way (Duan & Kollman 1998). Bernèche and Roux uses umbrella sampling to identify all the positions of the K+ ion along the KcsA K+ membrane channel. They show that the channel for K+ is virtually barrierless, hence it is a diffusion controlled process but more impressively, they identify two K+ sites just outside the channel, which was subsequently identified by electron density in a high-resolution structure (Bernèche & Roux 2001).

Rapid developments of technological know how in proteomics coupled with the improvement of bioinformatics tools have resulted in a deluge of structural information that guarantees acceleration in research. *In silico* identification of potential therapeutic targets in *Clostridium botulinum* by the approach subtractive genomics carried out Koteswara *et al.,* (2010). Evolutionarily related proteins have similar sequences and naturally occurring homologous proteins have similar protein structure. It has been shown that three-dimensional protein structure is evolutionarily more conserved than expected due to sequence conservation (Kaczanowski 2010).

Proteins fulfill several crucial functions, having catalytic, structural and regulatory roles in all organisms. Knowledge of the three dimentional structure of proteins is a basic prerequisite for understanding their function. It provides a basis for studies of substrate or ligand interactions with a particular enzyme or protein. Three dimensional structures of proteins are determined by X-ray crystallography and NMR spectroscopy. By 1st June 2007, more than 40000 sets of atomic coordinates for proteins have been deposited in the Protein Databank (Berman 2008).

The first modeling studies, carried out in the late 1960s ans early 1970s, wire based upon the construction of wire or plastic models.The later study are performed using interactive computer graphics. Browne et al. published the first report on homology modeling (Browne *et al.,* 1969). The model bovine α-lactalbumine on the known 3D structure of hen egg white lysozyme. Later on Warme et al. produced a model for α-lactalbumine on the basis of the crystal structure of lysozyme (Warme *et al.,* 1974). These model were constructed by taking the existing coordinates of the known structure, and mutating side chains not identical in the protein to be modeled. This approach to protein modeling is still employed today with considerable success, especially when the proteins are similar (May & Blundell 1994).

McLachlan and Shotton modeled α-lytic proteinase of the fungus Myxobacter 495 on the basis of the structures of the mammalian chymotrypsin and elastase (McLachlan & Shotton 1971). The modeling was a difficult task because the sequence identity between the protein to be modeled and the known structures was of the order of 18%. Subsequently, determined the crystal structure of α-lytic proteinase and compared the X-ray structure with the homology model. They found that although segment of both domains of the model were built correctly, misalignment of the sequence led to local errors.

Greer introduced the modeling of variable regions in proteins on the basis of equivalent region from homology proteins of known structures. In order to construct the homology models of a number of different serine proteases, he superimposed the structures of trypsin, chymotrypsin and elastase and found many equivalent Cα atoms within 1.0 A of one another (Greer 1980, 1981). The regions comprised of the amino acids of these Ca atoms were described as structurally conserved regions (SCRs). All of the remaining positions corresponded to structurally variable or loop regions (VR) where the insertions/deletions were located. The main chain of both structurally conserved (SCR) and variable regions (VR) was built from the fragments of known serine proteases. The side chains were modeled according to the conformation found at the equivalent positions for those identical side chains in the known structures.

Among the aspartic proteinases the first models were constructed for rennin and rennin inhibitor complexes using the 3D structure of the distantly related fungal proteinases (1985; Akahane *et al.,* 1985). Later on, the homology models for rennin were built using the structures of mammalian aspartic proteases, pepsin and chymosin (Fragao *et al.,* 1994; Hutchins and Greer 1991). Comparison of the rennin models constructed from fungal and mammalian enzyme revealed that errors in the models arose from the difference in the arrangement of helices and strands between the mammalian and fungal aspartic proteinases, as well as the rather different variable regions. Nevertheless, the active of rennin was modeled reasonably correctly.

In the early eightes, manual homology modeling was facilitated by manipulation of protein molecules on the graphics terminal that was made possible by computer programs such as FRODO (Jones 1978). Since the mid 1980s, a large number of homology models of proteins with different folds and functions have been reported in the literature (Johnson *et al.,* 1994; Sali 1995).

The sequence alignment and template structure are then used to produce a structural model of the target. Because protein structures are more conserved than DNA sequences, detectable levels of sequence similarity usually imply significant structural similarity (Marti-Renom *et al.,* 2000).

The quality of the homology model is dependent on the quality of the sequence alignment and template structure. The approach can be complicated by the presence of alignment gaps (commonly called indels) that indicate a structural region present in the target but not in the template, and by structure gaps in the template that arise from poor resolution in the experimental procedure (usually X-ray crystallography) used to solve the structure. Model quality declines with decreasing sequence identity; a typical model has ~1-2 Å root mean square deviation between the matched C^{α} atoms at 70% sequence identity but only 2-4 Å agreement at 25% sequence identity. However, the errors are significantly higher in the loop regions, where the amino acid sequences of the target and template proteins may be completely different.

Homology modeling can produce high-quality structural models when the target and template are closely related, which has inspired the formation of a structural genomics consortium dedicated to the production of representative experimental structures for all classes of protein folds (Williamson 2000). Like other methods of structure prediction, current practice in homology modeling is assessed in a biannual large-scale experiment known as the Critical Assessment of Techniques for Protein Structure Prediction, or CASP.

2.9: Future prospects

Food pathogenic bacteria research presents a completely new outlook in the post genomic era. The wealth of information obtained from the genome projects needs to be mined. As newer and newer toxins genes are discovered and sequenced novel insights are being gained. Bioinformatics in combination with metagenomics as well as metaproteomics approaches has the potential to give a more detailed picture that underlies pathogenicity as well as diseases. In absence of crysllographic or NMR structure Homology modeling will enlighten us about three dimensional structure and molecular dynamics simulation opening newer possibilities for exploring the molecular mechanism and activity of that toxin. At the end of the day scientific perception will continue to play a vital role in creating models that clarify the functions of pathogens in improved manner.

2.10: References

Aitken C & Cay E (1975) Clinical psychosomatic research". *International journal of psychiatry in medicine,* **6**: 29–41.

Akahane K Nakagava S & Umeyana H Moriguchi I, Hirose S, Iizuka K & Murakami K (1985) Three-dimensional structure of human renin *Hypertension* **7**:3-12

Baltrus DA, Amieva MR, Covacci A, Todd M. Lowe TM, Merrell DS, Ottemann KM, Stein M, Salama NR & Guillemin K (2009) The Complete Genome Sequence of *Helicobacter pylori* Strain G27 *Journal of Bacteriology,* 191:447-448

Berman HM (2008) The Protein Data Bank: a historical perspective. Acta Cryst. (2008). **A64**, 88–95

Benedito VA, Dai X, He J, Zhao XP & Udvardi MK (2006) Functional genomics of plant transporters in legume nodules. *Func Plant Biol* 33:731-736

Bennetzen JL & Hall BD (1982) Codon selection in yeast. *J. Biol. Chem.,* 257:3026–3031.

Bentivoglio M & Pacini P (1995) Filippo Pacini: a determined observer. *Brain Res. Bull.* **38**: 161–165.

Benzecri JP (1992) Correspondence analysis handbook. Marcel Dekker, New York

Bernèche S & Roux B (2001) Energetics of ion conduction through the K+ channel *Nature* **414**:73-77

Birney E, Clamp N & Hubbard T (2002) Databases and tools for browsing genomes. *Annu Rev Genomics Hum Genet* 3:293-310

Bloom M (2001) Biology in silico: The Bioinformatics revolution. *The American Biology Teacher* 63: 397-403

Borodovsky M, McIninch JD, Koonin EV, Rudd KE, Medigue C & Danchm A (1995) Detection of new genes in a bacterial genome using Markov models for 3 gene classes. *Nucleic Acids Res* **23**: 3554-3562

Borodovsky M, Rudd KE & Kounin EV (1994) Intrinsic and Extrinsic Approaches for Detecting Genes in a Bacterial Genome. *Nucleic Acids Res* **22**: 4756-4767

Brooks B & Karplus M (1983)Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor *PNAS* **80**:*6571-6575*

Browne WJ, North ACT, Philips DC, Brew K, Vanaman TC & Hill RC (1969) A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme, J. *Mol. Biol.,* **42**: 65-86

Burley, SK & Bonnano, JB (2002) Structuring the universe of proteins. *Annu Rev Genomics Hum Genet* **3**:243-262

Carbone A, Kepes F & Zinovyev A (2005) Codon bias signatures, organization of microorganisms in codon space and lifestyle. *Mol Biol Evol* **22**: 547-561

Case DA & Karplus M (1979) Dynamics of ligand binding to heme protein *J Mol Biol* **132**:*343-368*

Northrup SH, Pear MR, Lee CY, McCammon JA & Karplus M(1982) Dynamical theory of activated processes in globular proteins *PNAS* **82**:*4035-4039*

Chen SL, Lee W, Hottes AK, Shapiro L & McAdams HH (2004) Codon usage between genomes is constrained by genome-wide mutational processes. *Proc Natl Acad Sci USA* **101**: 3480-3485

Cowart GS & Thomason BM (1965) Immunofluorescent detection of Escherichia coli; incidence of certain serogroups suspected of being pathogenic. *Am. J. Dis. Child.* **110**: 131–136.

Drucker MM, Polliack A, Yeivin R & Sacks TG (1970) Immunofluorescent demonstration of enteropathogenic Escherichia coli in tissues of infants dying with enteritis". *Pediatrics* **46**: 855–864.

Duan Y& Kollman PA (1998) Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution *Science* **282***:*740-744

Eswarappa SM, Janice J, Nagarajan AG, Balasundaram SV, Karnam G, Dixit NM & Chakravortty D (2008) Differentially Evolved Genes of *Salmonella* Pathogenicity Islands: Insights into the Mechanism of Host Specificity in *Salmonella PLoS ONE.*3: e3829 doi: 10.1371/journal.pone.0003829.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF Kerlavage AR, Bult CJ, Kerlavage AR, Tomb JF, Dougherty BA, Merrick JM & *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512

Fragao C, Topham C, Dhanaraj V & Blundell TL (1994) Comparative modelling of human renin: A retrospective evaluation of the model with respect to the X-ray crystal structure *Pure Appl. Chem.,* **66**:43-50

Garcia E, Llull D, Lopez R (1999) Functional organization of the gene cluster involved in the synthesis of the pneumococcal capsule. *Int Microbiol.* 2:169–76.

Garcia E, Llull D, Lopez R (1999) Functional organization of the gene cluster involved in the synthesis of the pneumococcal capsule. *Int Microbiol.* **2**:169–76

GCG (1994) Programme Manual for the Wisconsin Package, Version 8, University of Wisconsin

Glantz PJ (1970) Unclassified Escherichia coli serogroup OXI isolated from fatal diarrhea of rabbits. *Can. J. Comp. Med.* **34**: 47–49.

Goldman N & Yang ZH (1994) Codon based model of nucleotide substitution for protein coding DNA sequences. *Mol Biol Evol* **11**: 725-736

Gouy M & Gautier C (1982) Codon usage in bacteria correlation with gene expressivity. *Nucleic Acids Res* **10**: 7055-7074

Grantham R, Gautier C & Gouy M (1980a) Codon frequencies in 119 genes confirm consistent choices of degenerate base according to genome type. *Nucleic Acids Res.,* **8**:892-1912

Grantham R, Gautier C, Gouy M, Jacobzone M & Mercier R (1981) Codon catalogue usage is a genome strategy for genome expressivity *Nucleic Acids Res.,* **9**: r43-r75

Grantham R, Gautier C, Gouy M, Mercier R & Pave A (1980b) Codon catalogue usage and the genome hypothesis. *Nucleic Acids Res* **8**: r49-r62

Greer J (1980) Model for haptoglobin heavy chain based upon structural homology. *Proc. Natl. Acad. Sc. USA,* **77**:3393-3397

Greer J (1981) Comparative Model-building of the Mammalian Serine Proteases *J. Mol.Biol.,* **153**:1027-1042

Gronroos JA (1957) Investigations on Escherichia coli O groups 1-25, 44 and 78 and serotypes 26:B6, 55:B5, 86:B7, 111:B4, 125:B15 and 126:B16; occurrence in faeces of healthy and diarrhoeal infants. *Annales medicinae experimentalis et biologiae Fenniae* **35**: 1–35.

Grosjean H & Fiers W (1982) Preferential codon usage in prokaryotic genes-the optimal codon anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* **18**: 199-209

Grote A, Hiller K, Scheer M, Munch R, Nortemann B, Hempel DC & Jahn D (2005) JCat: A novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res* **33**:W526-31

Guilbert C, Perahia D & Mouawad L (1995) A method to explore transition paths in macromolecules. Applications to hemoglobin and phosphoglycerate kinase *Computer Physics Communications* **91**:263-273

Gustafsson C, Govindarajan S & Minshull J (2004) Codon bias and heterologous protein expression. *Trends Biotechnol* **22**:346-353

Hentschel U& Hacker J (2001) Pathogenicity islands: the tip of the iceberg. *Microbes and Infection*, **3**:545–548

Howard-Jones N (1984) Robert Koch and the cholera vibrio: a centenary. *British medical journal (Clinical research ed.)* **288**: 379–381.

Hutchins C & Greer J (1991) Comparative modeling of proteins in the design of novel renin inhibitors. *Crit. Rev. Biochem. Mol. Biol.*, **26**:77-127

Ideker, T, Galitski, T & Hood, L (2001) A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* 2: 343-372

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons of its protein genes. *J Mol Biol* 146:1-21

Ikemura T (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. *J. Mol. Biol.*, **158**:573–597.

Ikemura T (1985) Codon usage and transfer-RNA content in unicellular and multicellular organisms. *Mol Biol Evol* 2:13-34

Israelachvili J, & Wennerström *H* (1996*)* Role of hydration and water structure in biological and colloidal interactions *Nature* **379***:21-225*

Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL (1994) Knowledge-based protein modeling, *Crit. Rev. Biochem. Mol. Biol.*, **29**: 1-68

T.A. Jones (1978) A graphics model building and refinement system for macromolecules, *J.Appl.* Cryst. **11**: 268-272.

Kaczanowski S & Zielenkiewicz P (2010). Why similar protein sequences encode similar threedimensional structures? Theoretical Chemistry Accounts **125**:543-50

Karlin S & Mrazek J (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol* **182**:5238-5250

Kellner R (1975) Psychotherapy in psychosomatic disorders". *Arch. Gen. Psychiatry* **32**: 1021–1028.

Kim HJ, Park SH, Lee TH, Nahm BH, Kim YR, Kim HY(2008) Microarray detection of food-borne pathogens using specific probes prepared by comparative genomics. *Biosens Bioelectron.* 24(2):238-246.

Koteswara Reddy G, Nagamalleswara Rao K, Phani Rama Krishna B & Aravind S(2010) *In silico* identification of potential therapeutic targets in *Clostridium botulinum* by the approach subtractive genomicsInternational *Journal of Bioinformatics Research*, 2:12-16

Krogh A, Mian IS & Haussler D (1994) A hidden Markov model that finds genes in *Escherichia coli* DNA. *Nucleic Acids Res* **22**: 4768-4778

Kurland CG (1991) Codon bias and gene-expression. *FEBS Letters* 285:165-169

Kyrpides NC (1999) Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects worldwide. *Bioinformatics* 15:773–774

Levitt M & Warshel A (1975) Computer simulation of protein folding *Nature* **253**:694-698

Linde K, Köditz H & Hasenjäger H (1966) Treatment of infantile diarrhea caused by pathogenic Escherichia coli strains with oral large doses of benzyl procaine penicillin (in Polish). *Pediatria polska* **41**: 905–12.

Lindsay JA, Ruzin A, Ross HF, Kurepina N & Novick RP (1998) The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in Staphylococcus aureus. *Mol Microbiol.* **29**:527–543.

Lloyd AT & Sharp PM (1992) CODONS - a microcomputer program for codon usage analysis. *J Heredity* **83**: 239-240

Lobry JR & Gautier C (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 *Escherichia coli* chromosome encoded genes. *Nucleic Acids Res* **22**: 3174-3180

M Levitt M & Sharon R(1988) Accurate simulation of protein dynamics in solution *PNAS* **85***:7557-7561*

Macqueen RL (1954) Isolation of Bact. coli O.26.B.6 from a child with recurrent diarrhea. *British medical journal* **1**:1475–1476.

Martin-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F & Sali A (2000) Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 29:291–325

May ACW & Blundell TL (1994) Automated comparative modelling of protein structures *Curr. Opin. Biotechnol* **5**:355-360

McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P,Courtney L, Porwollik S, Ali J, Dante M, Du F et al., (2001) Complete genome sequence of Salmonella enterica serovar Typhimurium LT2 *Nature.* **413**:852–856.

McClure WB (1955) A severe nursery epidemic of diarrhoea associated with Esch. coli type 111 B4. *Canadian Medical Association journal* **72**:83–88

McLachlan AD & Shotton DM Structural similarities between α-lytic protease of myxobacter 495 and elastase (1971) *Nature,* **229**:202-205

Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, Griffin PM & Tauxe RV(1999) Food-related illness and death in the United States. *Emerg Infect Dis.*,5:607-625.

Medigue C, Rouxel T, Vigier P, Henaut A & Danchin A (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* **222**: 851-856

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH & Teller E (1953) Equation of State Calculations by Fast Computing Machines *The Journal of Chemical Physics* **21**:1087-1092

Miller W, Kateryna DM, Nekrutenko A & Hardison RC (2004) Comparative genomics. *Annu Rev Genomics Hum Genet* 5:15-56

Nesti C, Poli G, Chicca M, Ambrosino P, Scapoli C Brrai I (1995) Phylogeny inferred from codon usage pattern in 31 organisms. *Comp Appli Biosciences* 11:167-171

Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, Churcher C, Mungall KL, Bentley SD, Holden MTG, Sebaihia M, Baker S, Basham D, Brooks K, Chillingworth T, Connerton P, Cronin A, Davis P, Davies RM, Dowd L, White N, Farrar J, Feltwell T, Hamlin N, Haque A, Hien TT, Holroyd S, Jagels K, Krogh A, Larsen TS, Leather S, Moule S, Ó'Gaora P, Parry C, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S & Barrell BG **(2001)** Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. *Nature.* **413**:848–852.

Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T & Holroyd S et al (2000) The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. *Nature,* **403**:665–668.

Paulley JW (1975) Cultural influences on the incidence and pattern of disease. *Psychotherapy and psychosomatics* **26**: 2–11.

Perez-Iratxeta, C, Andrade-Navarro, MA & Wren, JD (2006) Evolving research trends in bioinformatics. *Brief In Bioinform* **8**: 88-95

Post LE & Nomura M (1980) DNA sequences from the str operon of Escherichia coli. *J. Biol. Chem.,* **255:**4660–4666.

Pouwels PH & Leunissen JAM (1994) Divergence in codon usage of *Lactobacillus* species. *Nucleic Acids Res* **22**: 929-936

Puigbo P, Bravo IG & Vallve-Garcia S (2008) E-CAI: a novel server to estimate an expected value of codon adaptation index (e CAI). *BMC Bioinformatics* 9: 65 doi. 10.1186/1471-2105-9-65

Richardson SM, Wheelan SJ, Yarrington RM & Boeke JD (2006) GeneDesign: Rapid, automated design of multi kilobase synthetic genes. *Genome Res* **16**: 550-556

Rocha EP (2004) Codon usage bias from tRNA's point of view: redundancy, specialization and efficient decoding for translational optimization. *Genome Res* **14**: 2279-2286

Ruchman I & Dodd K (1947) The Isolation of a Strain of Escherichia coli Pathogenic for the Rabbit's Eye from a Patient with Diarrhea. *J. Bacteriol.* **53**:653–656.

Ryan KJ (2004) *Sherris Medical Microbiology, 4th ed.* McGraw Hill.

Sali A (1995) Modelling mutations and homologous proteins. *Curr. Opin. Biotechnol.* 6: 437-451

Schmidt W (1995) Phylogeny reconstruction for protein sequences based on amino acid properties. *J Mol Evol.* **41**: 522-530

Searls DB (2000) Bioinformatics tools for whole genomes. *Annu Rev Genomics Hum Genet* 1:251-279

Sharp PM, Stenico M, Peden JF & Lloyd AT (1993) Codon usage: Mutational bias, translational selection, or both? *Biochem Soc Trans* **21**:835- 841

Sharp PM, Tuohy TMF & Mosurski KR (1986) Codon usage in yeast cluster-analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* 14: 5125-5143

Sharp, PM & Li WH (1987) The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15:1281-1295

Smith HW & Gyles CL (1970) The relationship between two apparently different enterotoxins produced by enteropathogenic strains of Escherichia coli of porcine origin. *J. Med. Microbiol.* **3**: 387–401.

Supek F & Vlahovicek K (2004) INCA: Synonymous codon usage analysis and clustering by means of self-organizing map. *Bioinformatics* **20**: 2329-2330

Thioulouse J, Doledec S, Chessel D & Oliver JM (1995) ADE softeware: multivariate analysis and graphical display of environmental data. In *Sofware per l'Ambiente*, edited by Guariso G & Rizzoli A Patron editor, Bolonia.

Umashankar V, Arun Kumar V & Sudarsanam D (2007) ACUA: A software tool for automated codon usage analysis. *Bioinformation* **2**: 62-63

Villalobos A, Ness JE, Gustafsson C, Minshull J & Govindarajan S (2006) Gene designer: A synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics* **7**: 285

Wade N (2000) Genetic code of human life is cracked. *New York Times*, June 27.

Warme PK, Momany FA, Rumball SU, Tuttle RW & Sheraga HA (1974) Computation of structures of homologous proteins. alpha-lactalbumin from lysozyme *Biochemistry*, 13:768-782

William B (1979) *The History of Bacteriology*, Dover publications.

Williamson AR. (2000) Creating a structural genomics consortium. *Nat Struct Biol* 7 S1 (11s):953.

Wright F (1990) The effective number of codons used in a gene. *Gene* **87**: 23-29

Wu G, Bashir-Bello N & Freeland SJ 2006 The synthetic gene designer: A flexible web platform to explore sequence manipulation for heterologous expression. *Protein Expr Purif* **47**:441-445

Wu G, Culley DE & Zhang W (2005) Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. *Micobiology* **151**: 2175-2187

Chapter III

Comparative Analysis of Codon Usage Patterns and Identification of Predicted Highly expressed Genes in Five Salmonella Genomes

3.1: Introduction

Food-borne disease has been defined by the World Health Organization (WHO) as an ailment of transmittable or toxic nature caused by, or thought to be caused by, the consumption of food or water (Adams *et al.*, 2000). A number of bacteria are known to be linked with food-borne diseases. Prominent amongst them are *Salmonella, Shigella, Listeria,* Staphylococcus, *Vibrio,* etc. *Salmonella* is a gram-negative, motile, rod-shaped bacterial pathogen extensively occurring in animals, primarily in poultry and swine. Environmental sources of the bacterium throughout the world include water, soil, insects, factory surfaces, kitchen surfaces, animal faeces, raw meats, raw poultry, and raw sea foods (Wray & Sojka 1978) *Salmonella* causes substantial morbidity and mortality globally. The human-adapted serovars are responsible for typhoid, a systemic and life-threatening disease; whereas non-human-adapted serovars are normally accountable for gastroenteritis (Brown *et al.*, 2005).

The infection machinery of *Salmonella* involves a number of bacterial virulence genes, many of which are liable for invading, surviving, and replicating within host cells (Groisman & Ochman 1997). Recent work has exposed that a sizeable portion of *Salmonella typhimurium* genes are positioned in distinct chromosomal regions called pathogenicity islands (Groisman & Ochman 1997; Hacker *et al.*, 1997). The pathogenicity islands enclose genes associated with diseases and are often sources of toxins. Their G+C content differs from the rest of the chromosome, signifying that horizontal gene transfer acquired them (Blum *et al.*, 1994). Besides these, *vir* genes, *hrp* genes, invasions, *pip* genes, SPI, SOP, and toxin genes are also associated with pathogenicity.

Like other branches of biology, the study of pathogenic microorganisms has undergone a paradigm shift. The incredible deluge of information from genome-sequencing projects is revolutionizing the science of bacterial pathogenicity. The accessibility of the complete genome sequences of *Salmonella* provides a scope to undertake bioinformatics-based approaches focusing on synonymous codon usage and investigating the gene expression profile of the organism.

The non-random usages of synonymous codons are well accredited (Ikemura 1985). Synonymous codon usage is species specific and differs appreciably between the genes in the same organism (Banerjee *et al.,* 2004). Mutational pressure and natural selection operating at the level of translation are the primary reasons behind codon usage variation among the genes in different organisms (Sen *et al.,* 2008). Codon bias is quite high in the highly expressed genes compared to lowly expressed ones inside a genome (Sharp & Li 1986; Sharp & Li 1987; Lafay *et al.,* 2000; Dos Resi. *et al.,* 2003). The bias of highly expressed genes is influenced by translational selection; in contrast to lowly expressed genes, which is governed by mutational bias (Banerjee *et al.,* 2004). In order to inspect the patterns and cause of codon usage, many indices have been projected to assess the degree and direction of codon bias (Sharp & Li 198711). Amongst them, the codon adaptation index (CAI) was proposed as an estimate of codon usage within a gene relative to a reference set of genes (by and large, ribosomal protein genes) (Sharp & Li 1987). This index has been revealed to relate better with mRNA expression levels (Ikemura 1981). Over and above codon adaptation index, the effective number of codons (Nc), (Wright 1990) which is described as the amount of equal codons producing the same codon usage bias as observed; and the incidence of optimal codons (Fop), (Sen *et al.,* 2008) defined as the proportion of synonymous codons that are optimal codons, are also used.

The objective of this study was to execute a comparative analysis of the synonymous codon usage patterns, predict expression levels for the protein coding genes in these pathogenic bacteria with special reference to the genes linked with pathogenicity, examine horizontally transferred pathogenicity-related genes to detect their presence in the strains, and scrutinize the nature of highly expressed genes to infer upon their lifestyle. We consider that the result of this study would be helpful for the microbiologists working on this bacterium.

3.2: Materials and Methods

The complete genome sequences for five *Salmonella* strains [(*Salmonella enterica* Paratyphi, *Salmonella enterica* Typhi CT18, *Salmonella enterica* Typhi Ty2, *Salmonella enterica cholerasuis* SC-b67, and *Salmonella typhimurium* LT2 (hence forth, these strains will be referred to as SEP, SECT18, SETY2, SECSCB67, and STLT2 respectively)] were obtained from the IMG website (www.img.jgi.doe. gov) (Markowitz *et al.*, 2006). All of the protein coding genes, genes associated with pathogenicity, and ribosomal protein genes were examined using Codon W software (http://bioweb2.pasteur.fr) (Sen *et al.*, 2008) and CAI Calculator 2 (http://www.evolvingcode.net/codon/CalculateCAIs.php.) (Wu *et al.*, 2005).

The software Codon W (Sen *et al.*, 2008) was employed to inspect G or C in the third position of codons (GC3s), as well as to determine the effective number of codons (Nc) (Wright 1990) and the frequency of optimal codons (Fop) (Sen *et al.*, 2008). Nc is a straightforward measure of codon bias (Wu *et al.*, 2005). It ranges from 20 (when merely one codon is used per amino acid) to 61 (when each and every codon is used in equal likelihood). Fop (Sen *et al.*, 2008) determines the section of synonymous codons that are optimal codons. Its value varies from 0 (meaning a gene has no optimal codons) to 1.0 (when a gene is exclusively comprised of optimal codons).

The codon adaptation index (CAI) (Sen *et al.*, 2008) values were computed using the web-based application the CAI Calculator 2 (http://www.evolvingcode.net/codon/cai/ cais.php) (Wu *et al.*, 2005) taking the ribosomal genes as a reference. It quantifies the relative adaptiveness of a gene's codon usage, which is its codon usage as compared to the codon usage of highly expressed genes. The relative adaptiveness of each codon is the quantity of the usage of each codon compared to that of the most plentiful codon inside the same synonymous family (Sen *et al.,* 2008). The CAI value varies from 0 to 1.0, with higher CAI values signifying that the gene of concern has a codon usage pattern resembling that in the reference genes.

Z test was performed to check whether the values of the above-mentioned indices in the pathogenicity-related genes and ribosomal protein genes varied from those in the protein coding genes.

An analysis of the horizontally transferred pathogenicity-related genes among the studied strains was carried out to detect whether they are present in all the strains or native to a particular strain. The information about horizontally transferred genes was obtained from the website (http://cbcsrv.watson.ibm.com/HGT/). Tsirigos and Rigoutsos devised a new computational method for identifying horizontally transferred genes in 123 microbial genomes. It relied upon a gene's compositional features and necessitated having knowledge on codon boundaries. In addition to the single genes, the method was applicable to the clusters of genes transferred horizontally. The technique conveys a typicality score to each gene reflecting the gene's similarity with the containing genome, using specific features (Tsirigos & Rigoutsos 2005).

First of all, the pathogenicity-related genes acquired by horizontal gene transfer mechanisms in the studied strains were sorted out. Using the Integrated Microbial Genomes database (www.img.jgi.doe.gov.), (Markowitz *et al.*, 2006) the sorted pathogenicity-related genes for each strain were subjected to IMG Genome BLAST against the studied strains to find out the sequence homologs. The minimum percent identity was set at 90%; and the maximum E (expect) value 1e-2.

Correspondence analysis (COA) was performed using Codon W (http://bioweb2.pasteur.fr) (Sen *et al.,* 2008). This method explores the major trends in codon and amino acid variations among the genes.

3.3: Results

Codon usage patterns

Our first endeavour in the study of the codon usage patterns among various *Salmonella* genomes was to settle on the degree of variation in codon use. Most bacteria with a balanced AT/GC genome content have a sizeable amount of codon variation. Codon heterogeneity is usually associated with gene expression level. Thus, highly expressed genes contain a higher frequency of codons that are considered translationally optimal (Ikemura 1985; Lafay *et al.,*2000; Ikemura 19857).

The GC3s and Nc values for all of the genes in these genomes were calculated to determine if codon heterogeneity exists among genes of various *Salmonella* species. Two different indices, namely, effective number of codons (Nc) and GC3, have been used to detect codon usage variation among the genes. The Nc vs. GC3 plots have been recommended to be an effective means to explore the codon usage variations among genes in the same genome (Wright 1990). The Nc values of the *Salmonella* genes ranged from 25±1 to 61±0, and the GC3 values ranged from 0.14±0.3 to 0.91±0.6.

From Fig. 1, it is seen that the pathogenicity-related genes are lying below the expected curve. Genes which are anticipated to be highly expressed are clustered at one end of the Nc/GC3 plots. This phenomenon has been previously reported in *E. coli* and *Streptomyces* (Wu *et al.,* 2007).

Table 1 show that the mean Nc values of the total protein coding genes in the studied strains are in the range of 46-47, with the mean standard deviation value hovering around 6. With the exception of the ribosomal protein genes, the mean Nc values of the other categories of genes in the studied strains are quite high.

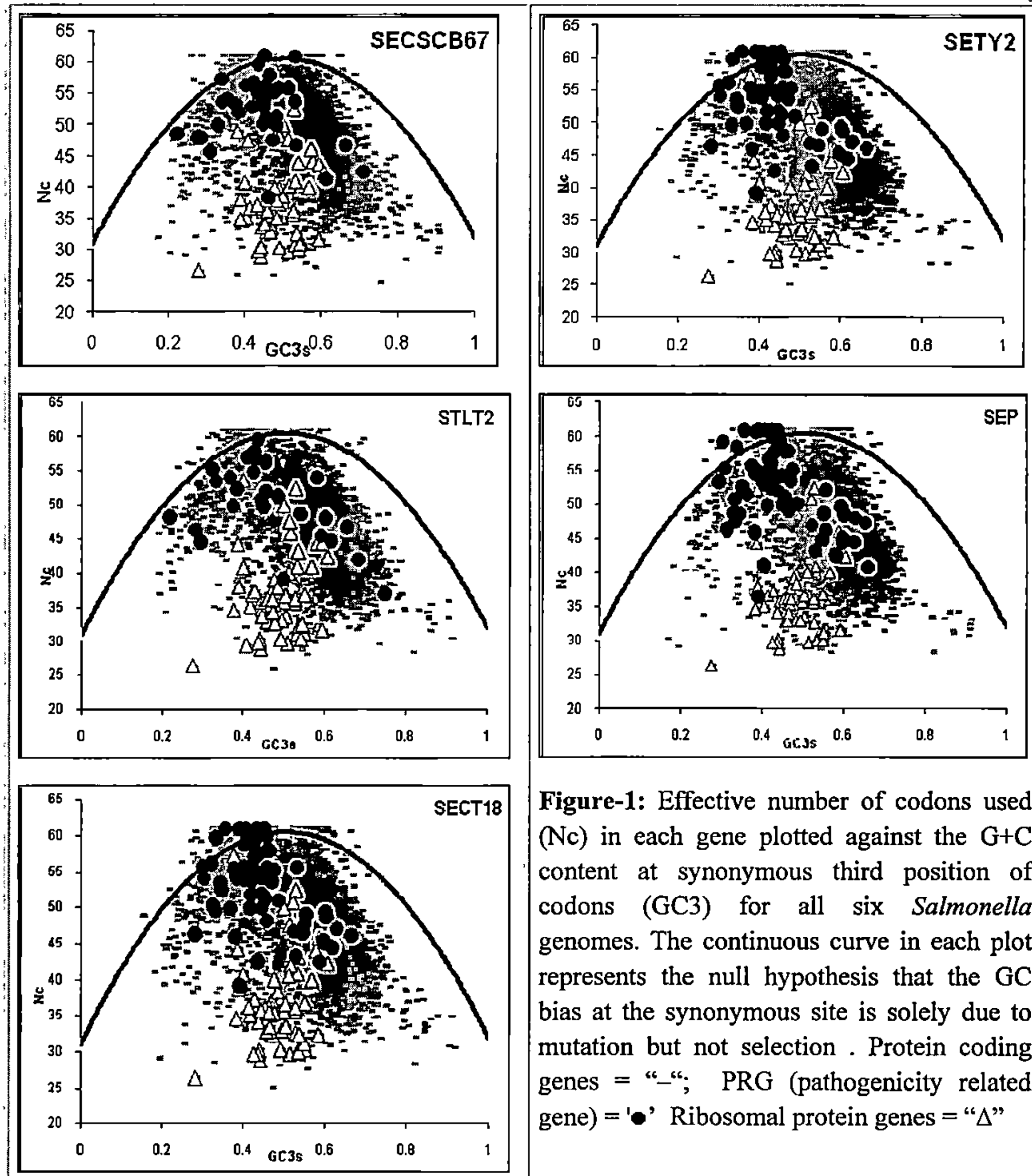


Figure-1: Effective number of codons used (Nc) in each gene plotted against the G+C content at synonymous third position of codons (GC3) for all six *Salmonella* genomes. The continuous curve in each plot represents the null hypothesis that the GC bias at the synonymous site is solely due to mutation but not selection . Protein coding genes = "–"; PRG (pathogenicity related gene) = '●' Ribosomal protein genes = "Δ"

Table 1: Mean values of effective number of codons (Nc), guanine cytosine percentage (GC), guanine cytosine ratio at third position (GC3), codon adaptation index (CAI), and frequency of optimal codons (Fop) of the genes in five *Salmonella* strains

Strain	**Genes**	**Nc**	**GC%**	**GC3%**	**CAI**	**Fop**
SEP	PCG	46.2±6.07	52.5±0.055	52.53±0.095	0.450±0.088	0.528±0.086
	RPG	36.90±6.05	51.20±0.02	48.6±0.062	0.705±0.102	0.738±0.090
	PRG	52.2±5.26	46.2±0.053	44.4±0.092	0.378±0.054	0.435±0.066
SECT18	PCG	47.26±6.28	52.1±0.056	55.4±0.098	0.445±0.086	0.522±0.085
	RPG	36.91±6.24	51.11±0.028	48.51±0.060	0.704±0.103	0.738±0.091
	PRG	52.61±5.73	46.32±0.050	44.30±0.087	0.377±0.055	0.434±0.064
SETY2	PCG	46.82±6.08	52.42±0.055	55.87±0.094	0.448±0.088	0.526±0.085
	RPG	36.91±6.24	51.17±0.028	48.57±0.061	0.704±0.103	0.738±0.103
	PRG	52.61±5.88	46.25±0.051	44.16±0.088	0.375±0.056	0.429±0.63
SECSCB67	PCG	47.12±6.22	52.31±0.05	55.64±0.096	0.460±0.083	0.522±0.084
	RPG	37.23±6.25	51.01±0.031	48.29±0.063	0.699±0.117	0.727±0.109
	PRG	52.08±5.17	45.67±0.056	44.31±0.098	0.412±0.037	0.449±0.055
STLT2	PCG	46.69±5.95	52.65±0.053	56.23±0.093	0.444±0.087	0.527±0.083
	RPG	36.52±5.92	51.19±0.029	48.55±0.062	0.710±0.100	0.743±0.089
	PRG	50.43±5.77	46.71±0.071	46.33±0.127	0.391±0.049	0.458±0.069

SEP= *Salmonella enterica* Paratyphi; SECT18= *Salmonella enterica* Typhi CT18; SETY2= *Salmonella enterica* Typhi Ty2; SECSCB67= *Salmonella enterica cholerasuis* SC-b67; STLT2= *Salmonella typhimurium* LT2; PCG= protein coding genes; RPG= ribosomal protein genes; PRG= pathogenicity-related genes

From Table 1 it is observed that there is a good deal of variation of GC3 values among different categories of genes in the studied strains. Variation in the mean Nc values and GC3 values for the different gene groups was observed within the same species as well as other species. Ribosomal protein genes and the protein coding genes had higher Fop values compared to the pathogenicity-related genes.

Z test did not reveal any significant difference between the different types of genes undertaken in the study at significance level of 0.05%. Z test gives a standard normal cumulative distribution function. For a given hypothesized population mean, Z test returns the probability that the sample mean would be greater than the average of observations in the data set (array) - that is, the observed sample mean.

Table 2: Two-tailed probability values of Z test for codon adaptation index values in pathogenicity-related genes, ribosomal protein genes, and protein coding genes

Strain	Genes	Probability values of z test	Correlation with *P* values
SEP	PCG	0.854	0.192
	RPG	0.455	0.079
	PRG	0.498	-0.105
SECT18	PCG	0.494	-0.005
	RPG	0.480	-0.128
	PRG	0.496	0.142
SETY2	PCG	0.500	0.193
	RPG	0.479	-0.094
	PRG	0.495	-0.240
SECSCB67	PCG	0.500	-0.08
	RPG	0.491	0.263
	PRG	0.473	0.170
STLT2	PCG	0.654	0.200
	RPG	0.494	0.165
	PRG	0.945	-0.132

SEP= *Salmonella enterica* Paratyphi; SECT18= *Salmonella enterica* Typhi CT18; SETY2= *Salmonella enterica* Typhi Ty2; SECSCB67= *Salmonella enterica cholerasuis* SC-b67; STLT2= *Salmonella typhimurium* LT2; PCG= protein coding genes; RPG= ribosomal protein genes; PRG= pathogenicity-related genes

From table 2 it is clearly seen that two-tailed probability values of Z test for CAI values in pathogenicity-related genes, ribosomal protein genes, and protein coding genes reveal trivial differences in SEP and STLT2 and are more or less same in SECSCB67, SETY2, and SECT18. There is no significant correlation between the *P* values of the different sets of genes. The correlations have been depicted in Table 2.

Analysis of horizontally transferred pathogenicity-related genes

From the web location (http://cbcsrv.watson.ibm.com/ HGT/),(Tsirigos & Rigoutsos 2005) it was observed that the studied *Salmonella* strains contained 616, 555, 562, 604, and 558 horizontally transferred genes for SECSCB67, SEP, SECT18, STLT2, and SETY2 respectively. Among these the numbers of pathogenicity-related genes were 15, 33, 29, 11, and 18 for SECSCB67, SEP, SECT18, STLT2, and SETY2.

IMG genome BLAST results revealed homologs having sequence identity with a number of similar proteins in other strains. In SECSCB67 the pathogenicity-related genes like putative shiga-like toxin A subunit, vir K, pathogenicity island-encoded protein SPI3, virulence gene, cytoplasmic cell invasion proteins, secreted proteins in SOP, and outer membrane-associated proteins found 18 horizontally transferred homologs (percent identity ranging from 95 to 100) in STLT2, SEP, SECT18, and SETY2.

In SEP pathogenicity-related genes like type III secreted protein effector, putative pathogenicity island proteins, putative pathogenicity island lipoproteins, putative pathogenicity island effector protein, outer membrane invasion protein, outer membrane virulence proteins, toxinlike proteins, putative vir K proteins, virulence proteins, cell adherence invasions, virulence-associated secretary proteins, pathogenicity island 1 effector proteins, and oxygen-regulated invasins had 52 horizontally transferred homologs (percent identity, 96-100) in SECT18; SELT2, SETY2, and SECSCB67.

The SECT18 pathogenicity-related genes like putative auto transporter virulence proteins, putative pathogenicity island protein, putative pathogenicity island lipoproteins, putative pathogenicity island effector protein, outer membrane invasion protein, outer membrane virulence proteins, virulence proteins, cell invasion proteins, pathogenicity island 1 and 2 effector protein, cell adherence protein, hypothetical proteins associated with virulence, and invasion-associated proteins found 51 horizontally transferred homologs (percent identity, 95-100) in SELT2, SEP, SETY2, and SECSCB67.

Among the pathogenicity-related genes of SELT2, putative shiga-like toxin A protein, pathogenicity island-encoded protein A, virulence protein PAGD precursor, virulence proteins, and invasion protein transcriptional activators found 16 horizontally transferred homologs (percent identity, 95-100) in SETY2, SEP, SECT18, and SECSCB67.

In SETY2, the pathogenicity-related genes like putative pertussis-like toxin subunit A, outer membrane invasion protein, putative pathogenicity island effector protein, putative pathogenicity island protein, putative auto transporter/virulence factor, virulence protein, hypothetical protein associated with virulence, and invasion-associated secreted protein had 35 horizontally transferred homologs (percent identity, 95-100) in SELT2, SEP, SECT18, and SECSCB67.

Correlating codon usage bias with tRNA content in Salmonella genomes

Eduardo Rocha (Rocha 2004) discussed the correlation between codon usage bias and tRNA content in bacterial genomes. The optimal generation times of the five studied *Salmonella* genomes were obtained from personal communications with Prof. J. Parkhill, Sanger Institute, Welcome Trust Genome Campus. The studied *Salmonella* genomes had an optimal generation time of 0.5 to 1 hour and could be regarded as fast growers on the basis of Rocha's (Rocha 2004) observations. He reported that fast growers have a median of 61 tRNA genes compared to 44 for slow growers, and the former tend to have stronger codon usage bias contrary to the latter. SECSCB67, STLT2, SETY2, SECT18, and SEP had 85, 86, 78, 80, and 82 tRNA genes respectively. The studied *Salmonella* strains had on an average 37 distinctive anticodon tRNA genes, i.e., they had more similar tRNAs.

Multivariate statistical analysis

Multivariate statistical analysis was performed to study the codon usage variation among the genes. Correspondence analyses of codon count of the protein coding genes, ribosomal protein genes, and pathogenicity-related genes for the *Salmonella* strains were performed.

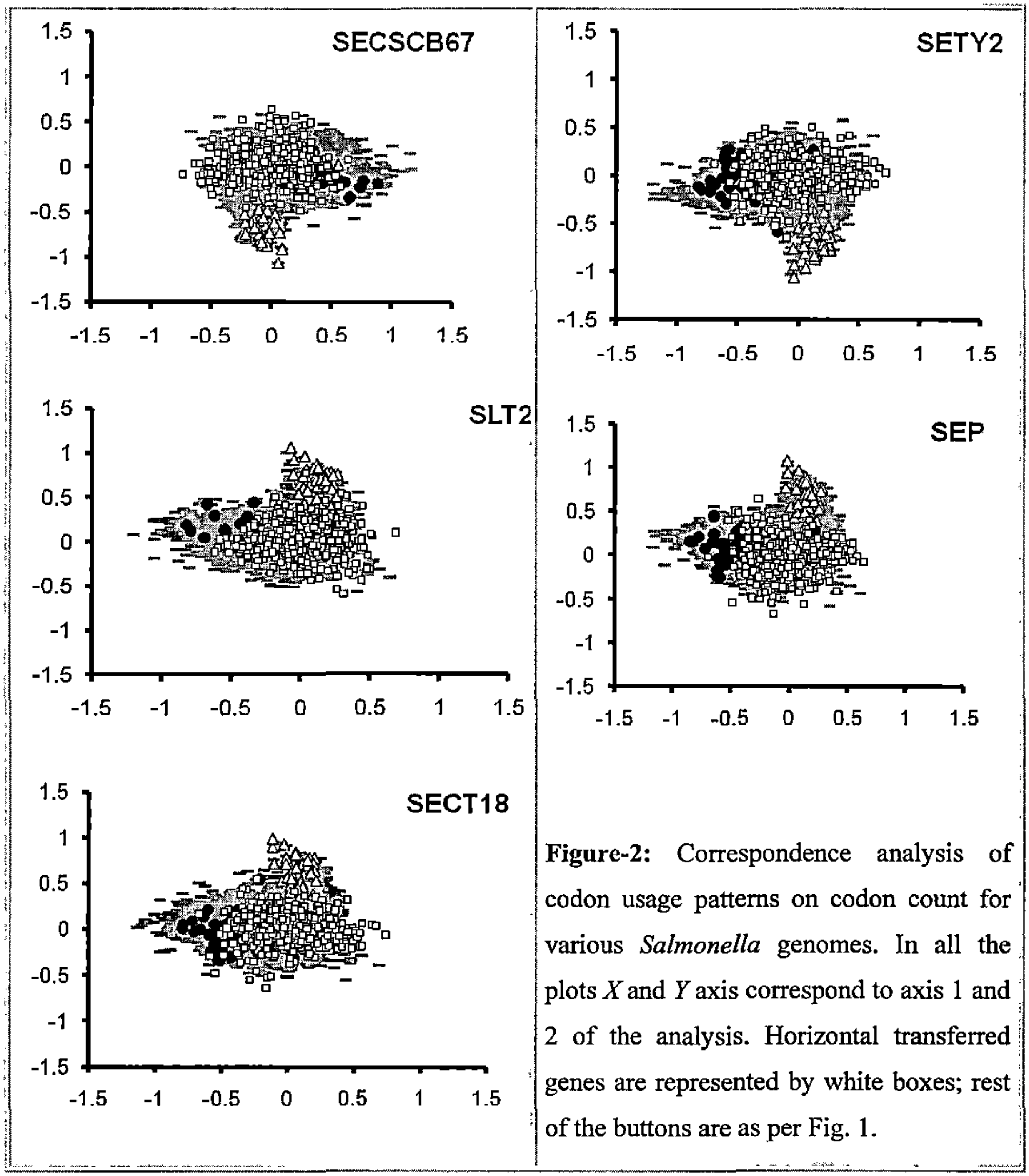


Figure-2: Correspondence analysis of codon usage patterns on codon count for various *Salmonella* genomes. In all the plots *X* and *Y* axis correspond to axis 1 and 2 of the analysis. Horizontal transferred genes are represented by white boxes; rest of the buttons are as per Fig. 1.

Fig. 2 reveals the positions of the genes on the planes defined by the first and second principal axes generated by COA of codon count for the protein coding genes, pathogenicity-related genes, and ribosomal protein genes. It is seen from Fig. 2 that the scatter plot of SEP, STLT2, and SECT18 revealed a small core region and two ascending horns, as reported for other eubacteria like *E. coli,* (Medigue *et al.,* 1991) whereas that of

SECSCB67 and SETY2 revealed a core region with two descending horns. Barring SECSCB67, the left horns in all the other genomes were less dispersed than the right horn. In case of SECSCB67, it is seen that the genes related to pathogenicity are located in the positive side of the Axis 1 compared to the same for other *Salmonella* genomes, where they reside on the negative side of Axis 1. Other genes are more or less clustered on the right side of the axis. The highly expressed genes are clustered together in the right horn of SEP, STLT2, SETY2, and SECT18 and left horn of SECSCB67 on the first axis of the COA of simple codon count.

No significant observation was noticed on correlating the CAI values of the protein coding genes of *Salmonella* strains with Axis 1. No correlation was observed between the positions of the genes on the Axis 1 produced by COA of codon count and the GC3 levels. However, we have found negative correlations between the positions of genes in Axis 1 produced by COA of codon count and Nc values of the protein coding genes in SECSCB67 and SECT18 and SETY2 (results not shown). Very little positive correlations were obtained between positions of genes in Axis 1 and Nc values in SEP and STLT2. The genes with negative coordinates on the principal axis have more biased usage of codons compared to the genes with positive Axis 1 coordinates.

Detection of PHX genes in Salmonella

Codon adaptation index (CAI) is a gauge of directional synonymous codon usage bias. The index uses a reference set of highly expressed genes from a species to evaluate the relative merits of each codon, and a score for a gene is determined from the frequency of use of all codons in that gene. The index assesses the degree to which selection has been successful in moulding the pattern of codon usage. The CAI value was calculated using the ribosomal protein genes, which are known to be highly expressed as a reference. The CAI values for all genes in different *Salmonella* strains were calculated, and their distributions are shown in Fig. 3.

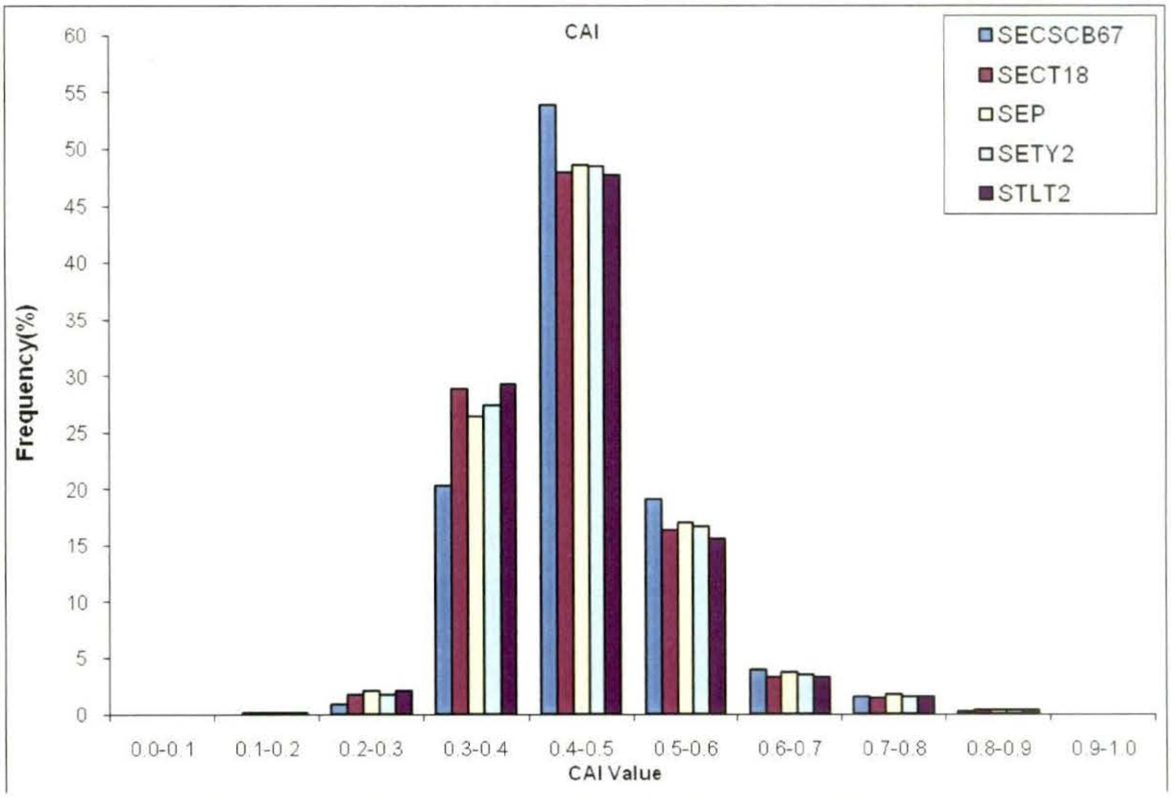


Figure-3 The frequency of distribution of the CAI values for all coding genes in the salmonella genomes

The average CAI values for different gene groups associated with diverse functions varied. Ribosomal protein genes showed high CAI values, indicating high levels of gene expression. These CAI values ranged from 0.203 to 0.877, 0.14 to 0.872, 0.191 to 0.874, 0.196 to 0.872, and 0.188 to 0.872 for SECSCB67, SECT18, SEP, SETY2, and STLT2 respectively. The majority of the genes for the *Salmonella* genomes had CAI values between 0.3 and 0.5

As visualized by Wu *et al.*,(2005) the top 10% of the genes, in terms of CAI values, were classified as the predicted highly expressed genes (PHX), and corresponded to CAI cutoffs of 0.562, 0.55, 0.558, 0.552, and 0.55 for SECSCB67, SECT18, SEP, SETY2, and STLT2 respectively. SECSCB67 had 477 PHX genes, including 51 ribosomal protein genes; SECT18 had 492 PHX genes, with 54 ribosomal protein genes; SEP had 423 PHX genes, with 54 ribosomal protein genes; SETY2 had 448 PHX genes, with 54 ribosomal protein genes; and SLT2 had 470 PHX genes, with 53 ribosomal protein genes.

Functional analysis of the PHX genes

To figure out the functional distribution of the PHX genes amongst the five *Salmonella* genomes, the clusters of orthologous groups of proteins were considered. For these *Salmonella* genomes, 20 COG categories were analysed.

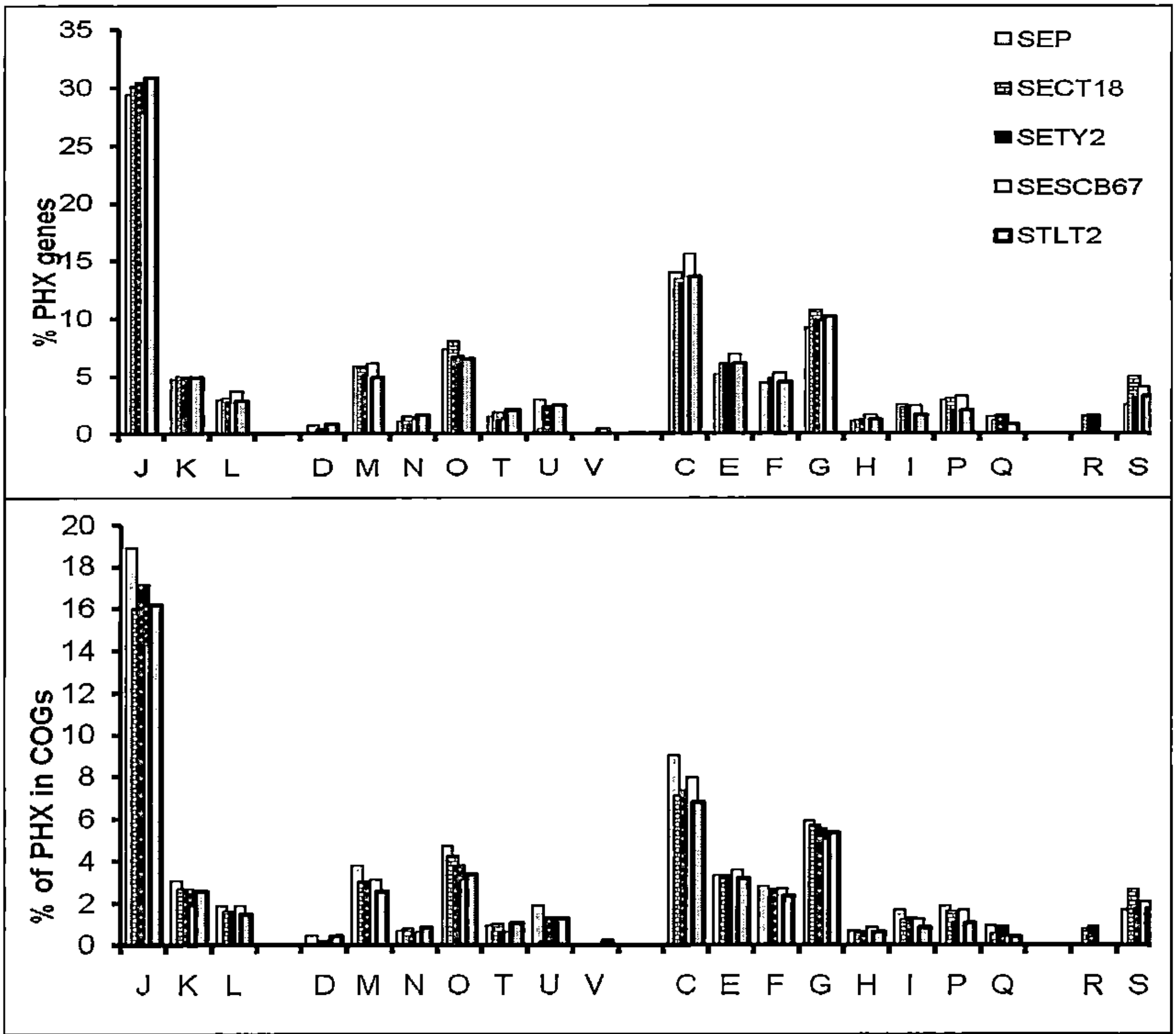


Figure-4: Distribution of Salmonella predicted highly expressed genes within functional COG groups (as in text)

Fig. 4 illustrates the allocation of the PHX into each COG category on the basis of total PHX genes (a) and the total genes within that COG group (b), expressed as a percentage.

To support the analysis, each of the COG categories were clustered in the following four COG groups: *information and storage processing* comprising of COGs connected to J-translation; K-transcription; L-DNA replication, recombination, repair (COG 1); *cellular processes* encompassing COGs linked to V-defence mechanism; T-signal transduction; M-cell envelope biogenesis; N-cell motility and secretion; U-intracellular trafficking; D-cell division; O-post-translational modification, protein turnover and chaperones (COG 2); *metabolism* consisting of COGs related to C-energy production and conversion; G-carbohydrate transport and metabolism; E-amino acid transport and metabolism; F-nucleotide transport and metabolism; H-coenzyme metabolism; P-inorganic ion transport and metabolism; I-lipid metabolism; Q-secondary metabolites, biosynthesis, and transport (COG 3); *general function prediction and unknown function* – R-general function prediction and S-unknown function (COG 4). The CAI values of each and every gene present in various COG groups were calculated, and the PHX genes were documented on the basis of the cut off values for various *Salmonella* genomes.

Fig. 4 exemplifies the percentage of PHX genes in different COG categories clustered in the four COG functional groups. The *Salmonella* genomes had the following distribution in the COG functional groups: SECSCB67 – 17.20, 9.10, 15.19, and 6.57%; SEP – 18.11, 9.72, 12.84, and 5.6%; SECT18 – 16.5, 10.20, 14.85, and 6.09%; SETY2 – 17.7, 9.48, 14.46, and 5.36; and STLT2 – 17.46, 9.15, 13.87, and 5.45 for the COG functional groups 1 to 4 correspondingly.

3.4: Discussion

The Nc and GC3 values for all genomes suggested that they exhibited differences in codon usage as anticipated. If synonymous codon bias were to be absolutely dictated by GC3s, Nc values should fall on the expected curve of the GC3 and Nc plot. However, we found that except for a few, the values obtained for majority of the genes were well below the expected curve (Fig. 1). This result clearly indicates that codon usage bias for the greater part of *Salmonella* genes is affected independently of overall base composition. On an average, the high Nc values of the protein coding genes and pathogenicity-related genes suggest that they are lowly biased. The clustering of highly expressed genes at one end of the Nc/GC3 plots in all the *Salmonella* genomes points out that codon usage in the studied *Salmonella* strains has a strong probability of being determined by translational selection.

On the whole, the GC3 content for these *Salmonella* genomes was moderate. Ribosomal protein genes and pathogenicity-related genes had lower GC3 values compared to the protein coding genes. Consequently, there are factors other than compositional constraints infl- uencing codon usage variation among the genes. Higher Fop values of the ribosomal protein genes and protein coding genes compared to pathogenicity-related genes imply the presence of higher proportion of optimal codons in these genes. If mutational bias had wholly controlled codon bias, these genes would have had a low Fop value. Since that was not the condition for these *Salmonella* genomes, there may be additional factors like gene expression levels and GC3 compositional bias acting on codon usage bias.

It is seen from the results of the Z scores in Table 2 that there is no significant correlation between the *P* values of the different categories of the genes in the studied genomes of *Salmonella.* So, the values for CAI in *Salmonella* genomes do not significantly differ in the categories of genes studied. These observations imply that there are inconsequential divergences in the characteristics of the studied genes.

The analysis of the pathogenicity-related genes revealed that not all of them were acquired by horizontal gene transfer mechanisms. Most of the pathogenicity-related genes acquired by horizontal gene transfer mechanisms were pathogenicity island encoded proteins, virulence proteins, secreted proteins, cell invasion proteins, toxin proteins, etc. Although the rest of the homologs for pathogenicity-related genes in all the strains showed percent identities ranging from 91 to 100, they were not found to be acquired by horizontal gene transfer mechanisms. These results indicated that they were native to those bacteria and they warded off the selective pressure of evolution. The horizontally transferred homologs, on the other hand, were gained from other organisms; and the high level of percent identity within the strains indicated that these genes are mobile within the genus. Most of them are associated with toxicity, virulence, pathogenicity islands, and invasion and are responsible for causing diseases resulting in epidemics. The high level of identity amongst them indicates that they evolved as a unit. Being a pathogenic bacterium, *Salmonella* has to fight against the host's defence systems, antibiotics, etc. The evolution of these genes as a unit suggests their ability to survive, infect, and exist as a pathogen.

Analysis of the correlation of codon usage bias with tRNA content in *Salmonella* genomes implies that these strains are well equipped to use small set of anticodons while maintaining high number of tRNAs. This is in line with Rocha's (Rocha 2004) observations. The ribosomal protein genes of these *Salmonella strains,* which are known to be highly expressed, showed high codon bias. This is expected since the codons associated with most abundant tRNAs have a propensity to be copious in highly expressed genes. The translation apparatus of *Salmonella* in all probability evolved with elevated codon bias in highly expressed genes compared to the rest of the genome. The mean CAI values of the studied *Salmonella* genomes varied widely from those of the ribosomal protein genes. This explains why selection for translational efficiency is the major source of variation of codon usage in *Salmonella* genomes. This has been previously exemplified by Rocha (Rocha 2004) in 102 bacterial genomes.

Multivariate statistical analysis data plotted in Fig. 2 specify that the relative positions of the pathogenicity-related genes and ribosomal protein genes are same in all the studied strains. It is fascinating to see that the highly expressed genes are clustered together in all the strains, signifying that they share a similar codon bias that is somewhat diverse from the rest of the genes. These results indicate that the translational selection is quite strong enough to ward off the selection pressure due to mutation in the studied strains of *Salmonella*. Majority of the genes in the core region (±0.5 to +0.5) are associated with housekeeping functions and metabolic pathways and are highly conserved. Genes located away from this core region included a number of hypothetical protein genes, ribosomal protein genes, and translation factors. In all the strains, the horizontally transferred genes were clustered together in the core region.

Absence of any significant correlation of the CAI values with Axis 1 of correspondence analysis of the protein coding genes of *Salmonella* strains clearly shows that expression levels do not discriminate genes according to their codon usage along the major explanatory axis. This was expected since the average CAI values of the protein coding genes are much lower than those of the ribosomal protein genes. In fact, a comparison of the results of different indices (Table 1) for ribosomal protein genes and all the protein coding genes reveals wide differences. These results validate our point that *Salmonella* genomes with lower mean CAI values are controlled by translational selection. No correlation of the positions of genes on the Axis 1 produced by COA of codon count with GC3 indicates that GC3 levels have practically no effect in differentiating the genes according to the codon usage variation along the first major explanatory axis. Negative correlation of the positions of genes in Axis 1 produced by COA of codon count with Nc values of the protein coding genes in SECSCB67 and SECT18 and SETY2 is attributed to the decrease in codon bias among the genes lying towards the left of Axis 1.

The plot of the frequency distribution of CAI values for the five *Salmonella* genomes showed more or less similar distribution patterns. All the genomes had a peak in the 0.4- 0.5

CAI range. CAI values for all the genomes rose and fell steadily. SEBSC67 had the highest peak value, viz., 53.90%. It has been noted that the percentages of PHX genes in COG category 1 and COG category 3 for the *Salmonella* genomes are well above the expected value of 10%. This reveals that the genes in these categories have reasonably superior expression levels than rest of the genes in the genomes. Functional analysis showed that the COG functional group 1 (information and storage processing) incorporated the maximum number of PHX genes in all the genomes. The COG groups translation, ribosomal structure biogenesis [J], and energy production and conversion [C] contained the highest number of predicted highly expressed genes. The distribution of high number of PHX genes in the translation, ribosomal structure biogenesis (J) functional groups of COG is attributed to the presence of high percentage of ribosomal protein genes which are highly expressed. Ribosomal protein genes which are PHX contributed to 67.94%, 66.66%, 67.08%, 70.42%, and 67.08% of PHX genes for SEP, STLT2, SETY2, SECSCB67, and SECT18 in the (J) functional group. Therefore, the weights of the ribosomal proteins played an important role in this case. Elevated number of PHX genes associated with translation, ribosomal structure biogenesis is beneficial for *Salmonella* to cause infections, overcome host immunity, and spread disease. The distribution patterns of the PHX genes in the various COG groups were approximately alike in all the five strains. Approximately 75% to 80% of the protein coding genes of the *Salmonella* strains belong to the COG category. This is significant because the huge number of genes in the COG groups of the *Salmonella* strains, in fact, helps them preserve their lifestyle, and it also divulges that *Salmonella* genomes are not subjected to genome reduction leading to gene loss. Being a pathogenic bacterium, it has to overcome host defence mechanisms to establish infection; and the presence of the genes responsible for pathogenicity and toxicity in the COG groups merely proves the fact.

The results from this study indicate variations existing among the genes of these genomes. Selection for translational efficiency is the major source of variation of codon usage in the genes of *Salmonella.* GC3 composition does not influence codon usage variation in the genes of these *Salmonella* strains. The horizontally transferred homologs, on the other hand, are gained from other organisms, and the high level of percent identity within the strains indicated that these genes are mobile within the genus. The evolution of these genes as a unit suggests their ability to survive, infect, and exist as a pathogen.

Correspondence analysis revealed clustering of the highly expressed genes together. Genes belonging to the COG categories are more or less conserved in the studied strains. Codon usage-based strategy has been applied to identify highly expressed genes in the studied strains of *Salmonella.* Genes related to information and storage processing include the highest number of PHX genes. Huge numbers of genes (approximately 75%-80%) in the COG categories of *Salmonella* genomes reflect their way of existence.

3.5: References

Adams MR & Moss MO (2000) Food microbiology. New Age International (P) Limited, Publishers

Banerjee T, Basak S, Gupta SK & Ghosh TC (2004) Evolutionary forces in shaping the codon and amino acid usages in *Blochmannia floridanus*. *J. Biomol. Struct. Dyn.*, **22**:13-23.

Blum G, Ott M, Lishewski A, Ritter A, Imrich H, Tschape H & Hacker J (1994) Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of *E. Coli* wild type pathogen. *Infect Immun.*, **62**:606-614.

Brown NF, Vallance BA, Coombes BK, Valdez Y, Coburn BA & Finley BB (2005) *Salmonella* pathogenicity island 2 Is expressed prior to penetrating the intestine. *PLoS Pathog* **1**:e32 DOI: 10.1371/journal.ppat.0010032

Dos Reis M, Wernisch L & Savva R (2003) Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucl. Acids Res.*, **31**:6976-6985.

Groisman EA & Ochman H (1997) How *Salmonella* became a pathogen. *Trends Microbiol.*, **5**:343-349.

Hacker J, Blum-Oheler G, Mulhdorfer I & Tschape H (1997) Pathogenicity islands of virulent bacteria: structure, function, impact on microbial evolution. *Mol. Microbiol.*, **23**:1089-1097.

Ikemura T 1985 Codon usage and transfer-RNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**:13-34.

Ikemura T (1981) Correlation between abundance of *E. coli* tRNAs and their occurrence of the respective codons in protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* system. *J. Mol. Biol.*, **146**:1-21.

Lafay B, Atherton JC & Sharp PM (2000) Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology,* **146**:851-860.

Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, Anderson I, Mavromatis K, Kunin V, Garcia Martin H, Dubchak I, Hugenholtz P & Kyrpides NC (2006) An experimental metagenome data management and analysis system. *Bioinformatics,* **22**:e359-367.

Medigue C, Viari A, Henaut A & Danchin A (1991) *Escherichia coli* molecular genetic map (1500 kbp) Update II. *Mol. Microbiol.,* **5**:2629-2640.

Rocha EP (2004) Codon usage bias from tRNA.s point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.***14**:2279-2286.

Sen A, Sur S, Bothra AK, Benson DR, Normand P &Tisa LS (2008). The implication of life style on codon usage patterns and predicted highly expressed genes for three *Frankia* Genomes. *Anton van Leeuw.*, **93**:335-346.

Sharp PM & Li WH (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.,* **24**:28-38.

Sharp PM & Li WH (1987) The codon adaptation index: A measure of directional synonymous codon usage bias, and its potential applications. *Nucl Acids Res.* **15**:1281-95.

Tsirigos A & Rigoutsos I (2005) A new computational method for the detection of horizontal gene transfer events. *Nucl Acids Res.* **33**:922-933.

Wray C & Sojka WJ (1978) Experimental *Salmonella typhimurium* infections in calves. *Res. Vet. Sci.,* **25**:139-43.

Wright F (1990) The □ □effective number of codons'' used in a gene. *Gene,* **87**:23-29

Wu G, Culley DE & Zhang W (2005) Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. *Microbiology,* **151:**2175-2187.

Chapter IV

Bioinformatic study of Pathogenicity related genes of three species of Helicobacter

4.1: Introduction

Isolated first by Marshall and Warren in 1982, from the gastric mucosa of a patient infected with gastritis and peptic ulceration, *Helicobacter pylori* represents a genus of gram negative, helix shaped, flagellated and motile, microaerophilic bacteria that persistently colonizes the human stomach (Warren & Marshall 1983). *Helicobacter pylori* have been linked with chronic antral gastritis and diseases like MALT lymphoma, peptic ulcer and gastric adenocarcinoma (Covacci *et al.,* 1999: Cover & Blaser 1999). More than 50% of the world's population harbor *H. pylori* in their upper gastrointestinal tract but merely 10% suffer from overt disease (Telford et al., 1994) Gastric cancer caused by H. pylori is a major cause of death worldwide particularly in the developing countries. The estimated current incidence of gastric cancer is approximately 16.2/100,000 persons per year (world standardized rate), with highest incidences in Eastern Asia, Eastern Europe, and South America (Ferlay *et al.,* 2004. The mechanisms of *H. pylori* transmission are incompletely characterized. Person-to-person transmission is most commonly implicated with fecal/oral, oral/oral, or gastric/oral pathways. The prevalence of H. pylori infection varies widely by geographic area, age, race, ethnicity, and SES. (Brown 2000)

The mechanisms by which bacterial infection leads to gastric mucosal damage include the direct effects of virulence factors produced by *H. pylori* like as *cag*A, *vac* A, or urease; the propagation and perpetuation of inflammation; oxidative stress; and the induction of apoptosis in infected gastric epithelial cells (Park & Hahm 2007). Virulent H. pylori strain genomes are characterized by the presence of a 40kb stretch or block of DNA containing nearly 31 open reading frames known as the cag pathogenicity island or cag PAI. The cag PAI has approximately 40 genes part of which codes for a complex type IV secretion system, which is able to translocate virulence proteins into the host cells. The most virulent gene, the cag A gene located within the PAI codes for a 1186 amino acid containing protein which confers H. pylori strains with the ability to cause ulcers.

Like other branches of biology, the study of pathogenic microorganisms has undergone a paradigm shift. The incredible deluge of information from genome sequencing projects is revolutionizing the science of bacterial pathogenicity. The accessibility of the complete genome sequences of *Helicobacter* provides a scope to undertake bioinformatics based approaches focusing on synonymous codon usage and investigating the gene expression profile of the organism.

The non-random usages of synonymous codons are well accredited (Ikemura 1985). Synonymous codon usage is species specific and differs appreciably between the genes in the same organism (Banerjee *et al.,* 2004). Unlike patterns of codon usage crop up from diverse factors. Mutational pressure and natural selection operating at the level of translation are the primary reasons behind codon usage variation among the genes in different organisms (Sen *et al.,* 2008). Codon bias is quite high in the highly expressed genes compared to lowly expressed ones inside a genome(Sharp & Li 1986; Sharp & Li 1987; Lafay *et al.,* 2000; Dos Reis 2003). The bias of highly expressed genes is influenced by translational selection in contrast to lowly expressed genes, which is governed by mutational bias (Banerjee *et al.,* 2004). In order to inspect the patterns and cause of codon usage, many indices have been projected to assess the degree and direction of codon bias Amongst them, the codon adaptation index (CAI) was proposed as an estimate of codon usage within a gene relative to a reference set of genes (by and large ribosomal protein genes) (Sharp & Li 1987). This index has been revealed to relate better with mRNA expression levels (Ikemura 1981). Over and above codon adaptation index, the effective number of codons (Nc) (Wright 1990), which is described as the amount of equal codons producing the same codon usage bias as observed, and the incidence of optimal codons

(Fop)(Sen *et al.,* 2008), defined as the proportion of synonymous codons that are optimal codons, are also used.

4.2: Materials and Methods

The complete genome sequences for three *Helicobacter* strains [*Helicobacter pylori* J99, *Helicobacter pylori* G27 & *Helicobacter pylori* P12 (hence forth, these strains will be referred to as HPJ, HPG & HP respectively) were obtained from the IMG website (www.img.jgi.doe.gov) (Markowitz 2006) Table 1 shows some of general features of those genomes. Table 2 shows Pathogenicity Related gene (PRG) in detail.

Table 1: Salient features of *Helicobacter* genomes analyzed in this study.

Organism	HPJ	HPG	HP
DNA, total number of bases	1643831	1663013	1684038
DNA G+C content (%)	39.19%	38.87%	38.79
Genes total number	1541	1547	1620
Protein coding genes	1494	1504	1578
RNA genes	47	43	42
rRNA genes	6	7	6
tRNA genes	36	36	36
Protein coding genes with function	897	1008	1128
Pseudogenes	3	0	0
Protein coding genes with enzymes	516	515	516
Protein coding genes with COGs	1087	1080	1082

HPJ: *Helicobacter pylori J99*, HPG: *Helicobacter pylori G27*, HP: *Helicobacter pylori P12*

Table-2: Pathogenicity Related gene (PRG) detail

Gene Name	HPJ	HPG	HP
Cag	15	20	22
adhesin	2	2	3
catalase	1	2	2
glr	1	1	1
trpA	1		1
trpB	1		1
urease	6	8	8
Virulence		1	2

All of the protein coding genes, genes associated with pathogenicity and ribosomal protein genes were examined using Codon W software (http://bioweb2.pasteur.fr) (sen *et al.,* 2008) and E-CAI server(http://genomes.urv.es/CAIcal/E-CAI.) (Puigbò et al., 2008).

The software Codon W (Sen *et al.,* 2008) was employed to inspect G or C in the third position of codons (GC3s), effective number of codons (Nc) (Wright 1990) and the frequency of optimal codons (Fop) (Sen *et al.,* 2008). The effective number of codons (Nc) is a straightforward measure of codon bias (Wu *et al.,* 2005). It ranges from 20 (when merely one codon is used per amino acid) to 61(when each and every codon is used in equal likelihood). Fop (Sen *et al.,* 2008) determines the section of synonymous codons that are optimal codons. Its value varies form 0 (meaning a gene has no optimal codons) to 1.0 (when a gene is exclusively comprised of optimal codons).

The 'codon adaptation index' (CAI) (Sen *et al.,* 2008) values were computed using a web-based application: 'E-CAI server' (http://genomes.urv.es/CAIcal/E-CAI.) (Puigbò et al., 2008) taking the codon usage table as a reference. It quantifies the relative adaptiveness of a gene's codon usage towards the codon usage of highly expressed genes. The relative adaptiveness of each codon is the quantity of the usage of each codon, to that of the most plentiful codon inside the same synonymous family (Sen *et al.,* 2008). The CAI value varies from 0 to 1.0 with higher CAI values signifying that the gene of concern has a codon usage pattern resembling that in the reference genes.

Using the Integrated Microbial Genomes database (www.img.jgi.doe.gov) (Markowitz *et al.,* 2006), the sorted pathogenicity related genes for each strain were subjected to IMG Genome BLAST against the studied strains to find out the sequence homologs. The minimum percent identity was set at 90% and the maximum E value 1e-2.

Correspondence analysis(COA) was performed usingCodonW (http://bioweb2.pasteur.fr)

(Sen *et al.,* 2008). This method explores the major trends in codon and amino acid variations among the genes

4.3: Results

The initial step in our work was to discriminate the three genomes of *Helicobacter* on the basis of their codon usage configuration. The Nc index of Wright (Wright 1990) is an appropriate tool to study codon degeneracy. It is a simple measure of overall codon bias and ranges from twenty to sixty one where 20 is the value obtained when only one codon is used for each amino acid (i.e., the codon bias is maximum) and 61 is the value obtained when all synonymous codon for each amino acid are equally used (i.e., no codon bias). The whole genome of H. pylori J99, H. pylori G27 and H. pylori P12 contains 1541, 1547 and 1620 genes respectively. The Nc score of all the genes in the genome of J99, G27 and P12 ranges between 22.56-61.0, 23.05-61.0 and 21.5-61.0 with a mean of 46.67, 46.27 and 46.33 respectively. The percentage of guanine and cytosine at the third position of a triplet, known as GC3 was also calculated alongside Nc to determine codon heterogeneity among the three species of H. pylori. Table 3 shows the mean values of different indices used to study codon usage patterns

Table 3: Mean values of Nc, GC, GC3, CAI , of the genes in three *Helicobacter* strains.

Organism	Genes	Mean Nc	Mean GC%	Mean GC3%	Mean CAI
HPJ	RPG	47.10926	41.65556	43.48333	0.736722
	PRG	49.25185	38.996	40.4963	0.711074
HPG	RPG	47.32857	41.783673	43.54693878	0.751734694
	PRG	48.01765	37.92941	38.44706	0.7485588
HP	RPG	46.792	41.712	43.172	0.7548
	PRG	47.595	37.7725	38.3075	0.748775

RPG = ribosomal protein genes; PRG = pathogenicity related gene;

Special emphasis was given to those genes of *H. pylori* which are an integral part of its genome and empowers *H. pylori* with pathogenic abilities. These genes include the cytotoxin associated pathogenic island genes commonly known as cag PAI (cagA, cagB, cagC, cagD, cagE, cagF, cagG, cagH, cagI, cagL, cagM, cagN, cagP, cagQ, cagS, cagT, cagU, cagV, cagW, cagX, cagy and cagZ) along with genes coding for a variety of adhesins, catalase, glutamate racemase, tryptophan synthase, urease and virulence factors. The cag pathogenicity island is represented by nearly 30 open reading frames and many of the genes are homologous to virulence associated genes of other bacterial pathogens. The cag PAI genes are also linked with amplified severity of disease in the human host (Eaton *et al.,* 2001). The mean Nc score of the genes responsible for pathogenicity in J99, G27 and P12 is 49.25, 48.01 and 47.59 respectively. Codon Adaptation Index or CAI was calculated for the entire genome of three strains of Helicobacter pylori. Codon Adaptation Index is a well-accepted parameter for studying the expressivity of a gene and assesses the extent to which the selection has been effective in moulding the pattern of codon usage. The Codon Adaptation Index ranges from 0 to 1.0, with higher CAI values signifying that the gene of concern has a higher degree of expressivity (Sharp & Li 1987). The CAI values ranges between 0.077 to 0.519 for J99, 0.075 to 0.419 for G27 and 0.077 to 0.454 for P12 respectively. Using ribosomal proteins of each genome of *Helicobacter* included in this study as standard for measuring gene expression levels it was observed that genes that may be considered as potentially highly expressive in all the three strains should have CAI values ranging from 0.65 to 0.8. Correspondence analyses of codon count of the protein coding genes, ribosomal protein genes and pathogenicity-related genes for the *Helicobacter* strains were performed and shown in Figure 1 which reveals the positions of the genes on the planes defined by the first and second principal axes generated by COA of codon count for the protein coding genes, pathogenicity-related genes, and ribosomal protein genes.

A multiple alignment of the pathogenicity related genes of the three species of *H. pylori* reveals that most of the *H. pylori* G27 pathogenicity genes such as cagA, cagD, cagF, cagG, cag M and others has significant identity with *H. pylori* P12 genes. Pair wise comparison of the pathogenicity genes of the three species based on the identity scores shows that the identity value mostly above 95%, the only exception being cagA gene. The identity score of *H pylori* P12 cagA gene with both *H. pylori* G27 and *H. pylori* J99 is significantly less at 87.

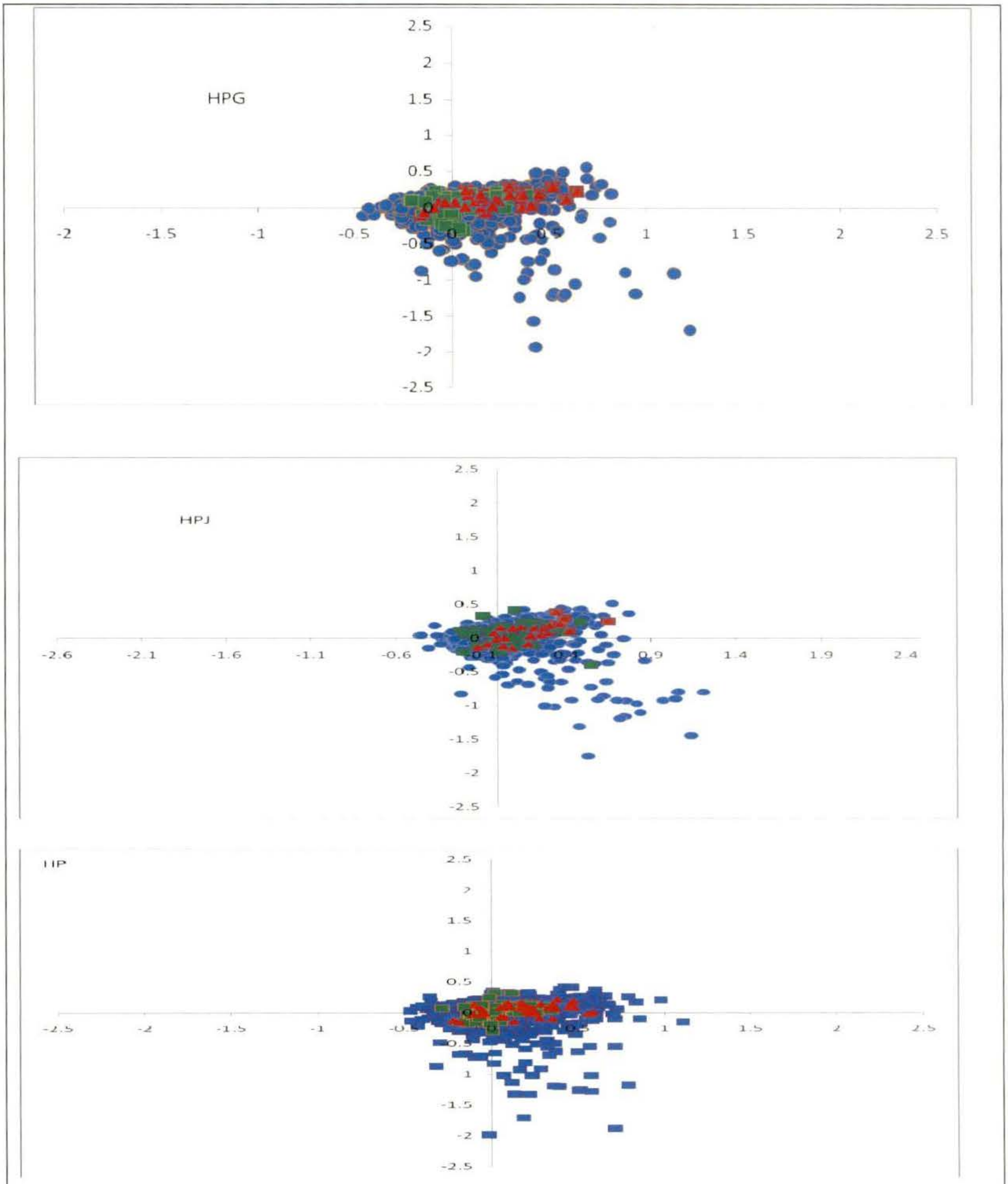


Figure 1: Correspondence analysis of codon usage patterns on codon count for *Helicobacter pylori* J99(HPJ), *Helicobacter pylori* G27(HPG) & *Helicobacter pylori* P12(HP). In all the plots *X* and *Y* axis correspond to axis 1 and 2 of the analysis. Protein coding genes = "blue" PRG (pathogenicity related gene) = "Red" Ribosomal protein genes = "Green"

4.4: Discussion

Detailed codon usage analysis of the cag PAI genes clearly demonstrates that cagD genes have a relatively higher Nc compared to the other cag PAI genes. The protein encoded by cagD gene is a covalent dimer which is an important component of the type IV secretion system (T4SS) that plays a crucial role in CagA translocation (Cendron *et al.,* 2008).

The most unpredictable observation was the apparent lack of genes with similar CAI values in the three *Helicobacter* strains. The cag PAI and other pathogenicity related genes of the three strains *of H. pylori* did not show a significant amount of anti-correlation of Nc with CAI as expected. But, it was quite remarkable to observe the fact that though the Nc values indicated lack of codon biasness, the CAI values pointed towards higher expression levels. This observation is quite significant taking into account the fact that codon biasness is one of the most, if not the most important factor in achieving higher gene expression levels.

The correspondence analysis of the three species of *Helicobacter pylori* indicates that in the case of *H. pylori* J99 the expression level of the cag and pathogenicity related genes is substantially similar to that of the ribosomal protein which clearly indicates that the pathogenicity related proteins are potentially highly expressed. Some sort of similar picture is also observed in the case of the other two species included in our study but in this case there are probably five pathogenicity related genes whose probable expression level does tally with that of any other ribosomal protein genes.

There is a high level of identity among the cagG, cagH and cage genes of the three species of *H. pylori.* The conservedness is consistent with their high CAI value, i.e., expression level and these genes have a crucial role stomach disease.

4.5: References

Banerjee T, Basak S, Gupta SK & Ghosh TC (2004) Evolutionary forces in shaping the codon and amino acid usages in *Blochmannia floridanus*. *J. Biomol. Struct. Dyn.*, **22**:13-23.

Brown LM (2000) Helicobacter pylori: epidemiology and routes of transmission. Epidemiol Rev.**22**:283-97

Cendron L, Couturier M, Angelini A, Barison N, Stein M & Zanotti G (2009) The *Helicobacter pylori* CagD (HP0545, Cag24) Protein Is Essential for CagA Translocation and Maximal Induction of Interleukin-8 Secretion Journal of Molecular Biology Volume 386:204-217

Covacci A, Telford JL, Giudice GD, Parsonnet J, & Rappuoli R (1999). *Helicobacter pylori* virulence and genetic geography. Science **284:**1328– 1333.

Cover TL & Blaser MJ (1999) *Helicobacter pylori* factors associated with disease. Gastroenterology **117:**257–261.

Dos Reis M, Wernisch L & Savva R (2003) Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucl. Acids Res.,* **31**:6976-85.

Eaton KA, Kersulyte D, Mefford M, Danon SJ, Krakowka S & Douglas E. Berg DE (2001) Role of *Helicobacter pylori cag* Region Genes in Colonization and Gastritis in Two Animal Models *Infection and Immunity*, 69:2902-2908

Ferlay J, Bray F, Pisani P & Parkin DM (2004) *GLOBOCAN 2002: Cancer Incidence, Mortality and Prevalence Worldwide.* IARC CancerBase No. 5, version 2 0, Lyon, France: IARC Press

Ikemura T (1981) Correlation between abundance of *E. coli* tRNAs and their occurrence of the respective codons in protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* system. *J. Mol. Biol.,* **146**:1-21.

Ikemura T (1985) Codon usage and transfer-RNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.,* **2**:13-34.

Lafay B, Atherton JC & Sharp PM (2000) Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori. Microbiology,* **146**:851-60.

Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, Anderson I, Mavromatis K, Kunin V, Garcia Martin H, Dubchak I, Hugenholtz P & Rocha EP (2004) Codon usage bias from tRNA.s point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.***14**:2279-86.

Park SW & Hahm KB(2007) Pathogenesis of Helicobacter pylori-induced Gastric Cell Injury *Hanyang Med Rev.* 27(3):34-40

Telford JL, Covacci A, Ghiara P, Montecucco C & R. Rappuoli R(1994) Unravelling the pathogenic role of *Helicobacter pylori* in peptic ulcer: potential new therapies and vaccines. Trends Biotechnol. **12:**420– 426.

Warren JR & Marshall B (1983) Unidentified curved bacilli on gastric epithelium in active chronic gastritis. Lancet. i:1273–1275.

Wright F (1990) The effective number of codons'' used in a gene. *Gene,* **87**:23-9

Wu G, Culley DE & Zhang W (2005) Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. *Microbiology,***151**: 2175-87.

Chapter V

Characterization of pathogenic genes through condensed matrix method, case study through bacterial zeta toxin

5.1: Introduction

The Zeta protein causes Gram-negative bacteria (Escherichia coli) to cease growing and form long cells with many chromosomes, clearly unable to divide. Similarly, when the toxin is introduced into yeast cells it changes their morphology and halts budding, and in large quantities it causes death. Human cancer cells also die from the toxic effect of the Zeta poison. The poison protein Zeta is extraordinarily large compared other toxins (287 vs. around 100 amino acids) and with the exception of its frequently encountered nucleotide binding motif, it does not show any similarities to any known proteins. The molecular mechanism by which the Zeta toxin operates has not yet been discovered (Zielenkiewcz & Dmowski 2009). Where alignment based and structural based phylogeny fails nucleotide triplet based method give light towards molecular phylogeny.

Availability of nucleotide sequences of zeta toxin motivated us to construct phylogram using their nucleotide sequence, which will complement the phylogram obtained by sequence similarity. In this work we have done the molecular phylogeny of zeta toxin using their nucleotide sequence and without making any sequence alignment. It is a based on a method developed by basak et al associating DNA sequences with a set of sequence invariant. In this work we have quantified the string, which favors the direct comparison of the sequences. A sequence invariant, as considered as a number independent of the labels A, C, G, T standing for adenine (A), cytosine (C), guanine (G), and thiamine (T). We have form the matrix associated with each sequence and calculated the leading Eigen value of the matrices to see the variation of leading Eigen values associated with the string and the relationship between the enzymes. We have also build a phylogram using the Eigen values of the characteristics matrices of zeta toxin.

Our results complement the observation with the earlier studies based on multiple sequence alignment and structural alignment. The uniqueness of this method is that it does not employ sequence alignment of complete nucleotide sequence of the corresponding gene.

5.2: Material and Methods

The nucleotide sequences of zeta toxin of some pathogenic and non pathogenic bacteria were obtained from www.img.jgi.doe.gov. In a DNA sequence of four letters, there are 64 possible triplets (subsequence of length3) that can occur, starting from AAA, AAT, AAG, AAC, ATA, ATT, ATG, ATC, AGA, AGT, AGG, AGC, ACA, ACT, ACG, ACC, etc. A 4×4×4 cubic matrix with 64 entries that denote the frequencies of occurrence of all the 64 triplets in a DNA sequence are introduced. For the cubic matrix, three groups of 4×4×4 matrices, {M1, M2, M3, M4}, {M5, M6, M7, M8}, {M9, M10, M11, M12}, can be obtained, each group of which contain all entries of the cubic (see Table I). Usually the group of 4×4 matrices {M1, M2, M3, M4} as the representative of the cubic matrix. The four matrices contain not only the information about frequencies of occurrence of all triplets of a DNA sequence but also the information about the frequencies of occurrence of pairs and every letter in a DNA sequence. For example, the number of all TG-pair in a DNA sequence is equal to the row sum of the third row in M2 plus ∂, where $\partial = 0$ if the last two letters of the DNA sequence are not TG and $\partial = 1$ otherwise. The frequency of occurrence of any pair in a DNA sequence can obtain by the above method. In addition, the frequencies of occurrence of four letters A, T, G, C are, respectively, equal to the sum of all entries of M1, M2, M3, M4 plus ∂, where ∂ are, respectively, equal to the number of A, T, G, C in the last two letters of the DNA sequence. The column sums of M1, M2, M3, M4 just denote the number of pairs of distance two in a DNA sequence (Randic, et al 2001; Randic & Basak 2001).

Table I: Three Groups of Four 4 × 4 matrices, {M_1, M_2, M_3, M_4},{M_5, M_6, M_7, M_8}, and {M_9, M_{10}, M_{11}, M_{12}}Listing All 64 possible XYZ Entries, Where X, Y, Z = A, C, G, T.

M_1				M_2				M_3				M_4			
AAA	AAT	AAG	AAC	TAA	TAT	TAG	TAC	GAA	GAT	GAG	GAC	CAA	CAT	CAG	CAC
ATA	ATT	ATG	ATC	TTA	TTT	TTG	TTC	GTA	GTT	GTG	GTC	CTA	CTT	CTG	CTC
AGA	AGT	AGG	AGC	TGA	TGT	TGG	TGC	GGA	GGT	GGG	GGC	CGA	CGT	CGG	CGC
AGA	ACT	ACG	ACC	TCA	TCT	TCG	TCC	GCA	GCT	GCG	GCC	CCA	CCT	CCG	CCC

M_5				M_6				M_7				M_8			
AAA	AAT	AAG	AAC	ATA	ATT	ATG	ATC	AGA	AGT	AGG	AGC	ACA	ACT	ACG	ACC
TAA	TAT	TAG	TAC	TTA	TTT	TTG	TTC	TGA	TGT	TGG	TGC	TCA	TCT	TCG	TCC
GAA	GAT	GAG	GAC	GTA	GTT	GTG	GTC	GGA	GGT	GGG	GGC	GCA	GCT	GCG	GCC
CAA	CAT	CAG	CAC	CTA	CTT	CTG	CTC	CGA	CGT	CGG	CGC	CCA	CCT	CCG	CCC

M_9				M_{10}				M_{11}				M_{12}			
AAA	TAA	AAG	CAA	AAT	TAT	GAT	CAG	AAG	TAG	GAG	CAG	AAC	TAC	GAC	CAC
ATA	TTA	GTA	TAC	ATT	TTT	GTT	CTG	TAG	TTG	GTG	CTG	ATC	CTT	TCG	CTC
AGA	TGA	GGA	CGA	AGT	TGT	GGT	CGG	AGG	TGG	GGG	CGG	AGC	GGC	GCG	CGC
ACA	TCA	GCA	ACC	ACT	TCT	GCT	CCG	ACG	TCG	GCG	CCG	ACC	CCT	GCC	CCC

We developed our own program in C++ to count all the possible triplets of t-RNA synthetase and formed the matrices by using all the possible triplets. Also we have calculated the leading Eigen values of the matrices by using MATHLAB (Version 4) (Toh *et al.,* 1999) software. We have constructed a distance matrix of the synthetases by summing the square of the difference of eigen values. A phylogram of the synthetases are constructed by the cluster analysis of the similarity matrix using phylip (Felsenstein 1989).

5.3: Results

The lengths of the zeta toxin of some pathogenic and non-pathogenic bacteria are given in Table 2. It is clear that the enzymes differ considerably in length. Firstly, we took the nucleotide sequence of zeta toxin of pathogenic and non pathogenic bacteria listed in Table 3 and counted the frequencies of occurrence of all the 64 triplets then the group of 4 × 4 matrices {M1, M2, M3, M4} as the representative of the cubic matrix are constructed.

Table 2: The lengths of the zeta toxin of some pathogenic and non-pathogenic bacteria

Bacteria Name	Short name	Length	Nature
Pseudomonas fluorescens	PSE	759	Non pathogenic
Frankia sp. CcI3	FR1	1053	Non pathogenic
Frankia sp. CcI3	FR2	1353	Non pathogenic
Mesorhizobium sp. BNC1 plasmid 1	MES	1761	Non pathogenic
Alteromonas macleodii 'Deep ecotype	ALT	720	Non pathogenic
Streptococcus pneumoniae ATCC 700669	STR	759	Pathogenic
Neisseria cinerea ATCC 14685	NEI	720	Pathogenic
Enterococcus faecalis TX0104	ENT	441	Pathogenic
Oribacterium sinus F0268	ORI	783	Pathogenic
O.algarvensis Gamma1	OLA	306	Non pathogenic
Crenothrix polyspora	CRE	2154	Non pathogenic

The leading Eigen values of each matrix are evaluated. The leading Eigen values of each matrix of those bacteria are represented in Table 4. The distance matrices of the synthetases are constructed by summing up the square of the difference of eigen values. The distance matrix for bacteria mentioned in Table 1 is given in Table 4.

Table 3: The leading Eigen values of each matrix of bacteria are represented in Table II.

Name of Bacteria	Short Name	M1	M2	M3	M4
Pseudomonas fluorescens	PSE	63.9742	59.4986	66.1942	61.1813
Frankia sp. CcI3	FR1	40.9835	45.5497	83.7531	1.0101
Frankia sp. CcI3	FR2	51.1907	38.5775	82.8222	1.001
Mesorhizobium sp. BNC1 plasmid 1	MES	39.4057	47.476	87.3356	86.5084
Alteromonas macleodii 'Deep ecotype	ALT	80.975	64.4988	59.6351	51.9077
Streptococcus pneumoniae ATCC 700669	STR	101.2985	68.0832	59.42	49.6625
Neisseria cinerea ATCC 14685	NEI	92.048	69.8199	70.3941	41.2137
Enterococcus faecalis TX0104	ENT	93.5377	88.5706	54.7326	37.3998
Oribacterium sinus F0268	ORI	108.5584	71.2367	65.8962	40.403
Olavius algarvensis Gamma1	OLA	71.8841	53.1558	57.2465	69.9674
Crenothrix polyspora	CRE	72.5779	69.704	60.1925	54.6687

Table 4: The distance matrix of zeta toxins of some pathogenic and non-pathogenic bacteria mentioned in Table 1

	PSE	FR1	FR2	MES	ALT	STR	NEI	ENT	ORI	OLA	CRE
PSE	0	4652.03	4499.27	1836.57	443.05	1645.37	1311.01	2416.11	2557.36	260.05	256.61
FR1	4652.03	0	153.67	7328.99	5130.63	7104.81	4991.42	6779.14	7096.86	6470.41	5015.98
FR2	4499.27	153.67	0	7549.96	4688.15	6296.98	4416.93	6406.48	6196.68	6051.22	4818.6
MES	1836.57	7328.99	7549.96	0	3982.31	6392.28	5609.09	8093.65	7932.02	2266.07	3344.99
ALT	443.05	5130.63	4688.15	3982.31	0	430.98	381.04	971.79	977.8	543.17	105.54
STR	1645.37	7104.81	6296.98	6392.28	430.98	0	280.4	652.31	190.33	1505.05	853.16
NEI	1311.01	4991.42	4416.93	5609.09	381.04	280.4	0	613.64	295.49	1683.91	664.21
ENT	2416.11	6779.14	6406.48	8093.65	971.79	652.31	613.64	0	659.73	2790.05	1123.29
ORI	2557.36	7096.86	6196.68	7932.02	977.8	190.33	295.49	659.73	0	2620.79	1532.99
OLA	260.05	6470.41	6051.22	2266.07	543.17	1505.05	1683.91	2790.05	2620.79	0	517.05
CRE	256.61	5015.98	4818.6	3344.99	105.54	853.16	664.21	1123.29	1532.99	517.05	0

Using the distance matrices phylograms are constructed, which are represented in Figures 1. Figure 2 represented the Phylogram of zeta toxins based on ClustalW (Thompson *et al.,* 1994).

Figure 1: Phylogram of zeta toxins based on condensed matrix method.

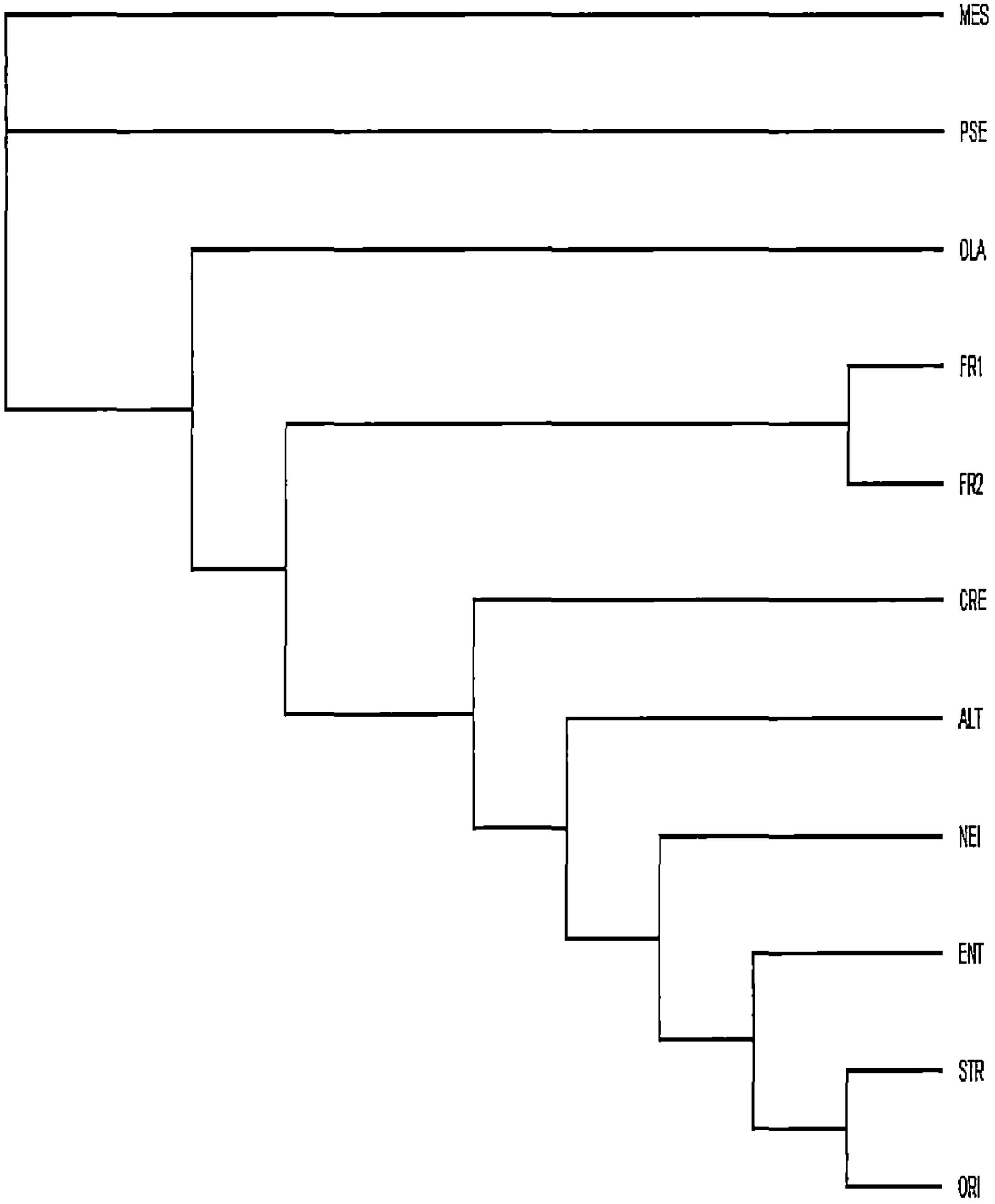
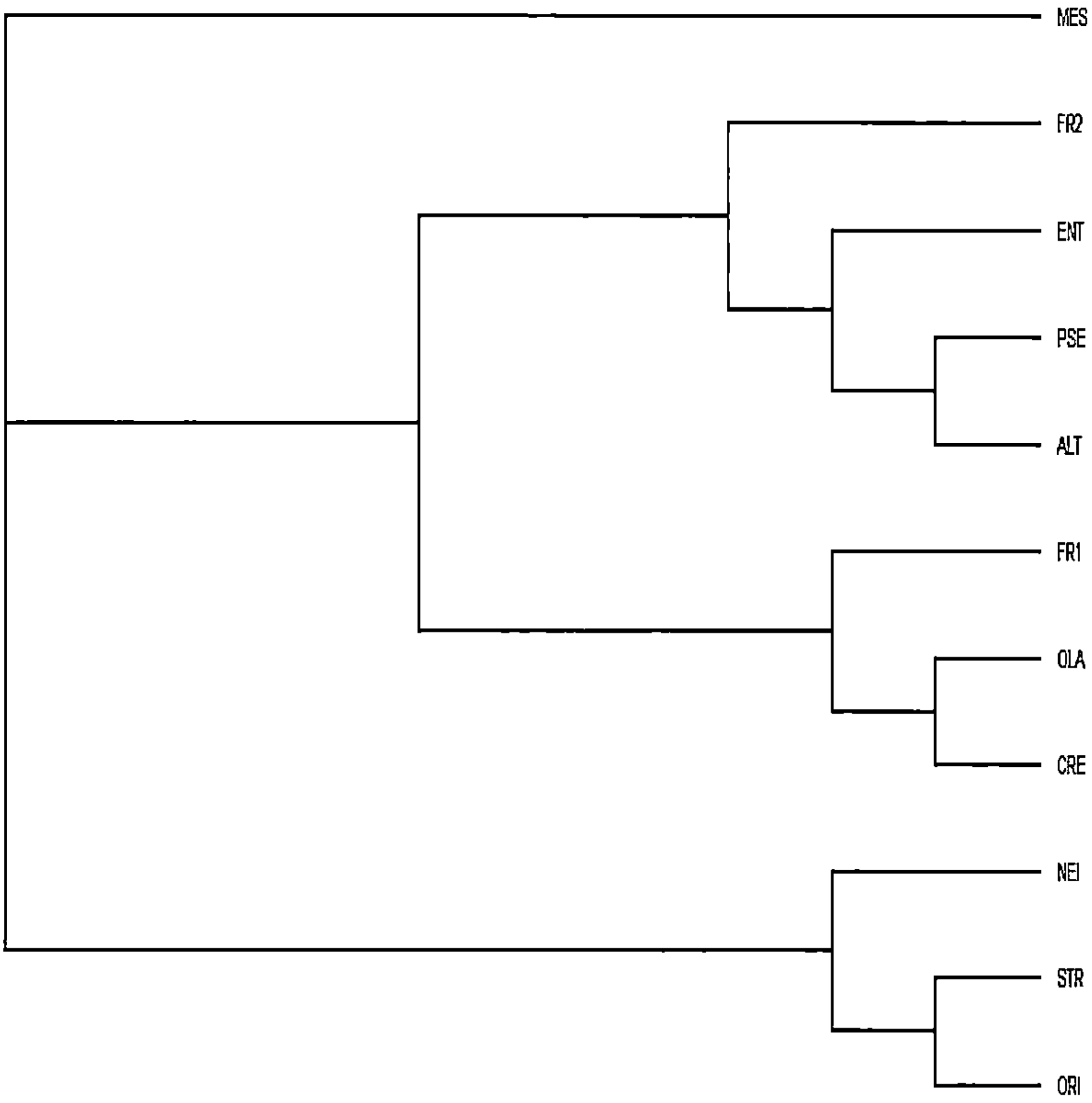


Figure 2: Phylogram of zeta toxins based on ClustalW.



5.4: Discussion

From Figure 1 is seen that, zeta toxin pathogenic and non-pathogenic bacteria forms different cluster. It is also seen that two zeta toxins of Frankia sp. Cc13 are in the same clade. From Figure 2 it is seen that separation are not clear between pathogenic and non-pathogenic bacteria which indicates superiority of condensed matrix method. It is also seen that two zeta toxins of Frankia sp. Cc13 are in the different clade.

Sequence comparison quickly becomes unreliable at this and lower levels of sequence identity. In this regime of similarity, it becomes difficult to distinguish between correctly aligned homologous sequences and unrelated sequences or random alignments. Structure based phylogeny has limited scope because adequate number of structures are not yet solved to draw any general conclusion. The nucleotide triplet based phylogeny is free from above mentioned limitations and it considers the full length of the genes for construction of phylogram. From the separation of zeta toxin by condensed matrix method may help to identify pathogenic or non-pathogenic strain/species.

5.3: References

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2), *Cladistics*, **5**:164-166.

Randic M & Basak SC (2001) Characterization of DNA Primary Sequences Based on the Average Distances between Bases, *J Chem Inf Comput Sci.*, **41**:561-568.

Randic M, Guo X & Basak SC (2001), On the characterization of DNA primary sequences by triplet of nucleic Acid Based, *J Chem Inf Comput Sci.*, **41**: 619-626.

Thompson JD, Higgins DG & Gibson TJ (1994), CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673-4680

Toh KC, Todd MJ & Tutuncu RH (1999) SDPT3 - a Matlab software package for semidefinite programming, Optimization Methods and Software, **11**:545--581.

Zielenkiewcz U & Dmowski M (2009) Systems ensuring plasmid inheritance Epsilon vs. Zeta. *Research in Progress Microbiology*, **1**: 38-39.

Chapter VI

Molecular Dynamics Simulation Receptor-Binding C-Terminal Domain from Clostridium difficile Toxin A

6.1: Introduction

Gastrointestinal diseases like antibiotic-associated diarrhea are caused by the gram positive bacteria like *Clostridium difficile* (Bartlett *et al.,* 1978; George *et al.,* 1978; Bartlett *et al.,* 1980a; Kelly & LaMont 1998). A strong correlation between pseudomembranous colitis, antibiotic therapy, *C. difficile* colonization and cytotoxin production are reported by various researchers (Aronsson *et al.,* 1981; Bartlett *et al.,* 1980b; Burdon *et al.,* 1981; Don & Devis 1981; George *et al.,* 1982; Meyers *et al.,* 1981; Tedesco, 1981; Willey & Bartlett 1979). Protective antibodies produced by the host immune system against the toxins of *C. difficile* play an important role in reducing the severity of disease. Despite the lack of understanding of the epidemiology of *C. difficile* disease, there has been a considerable progress towards the understanding of organism's mechanism of virulence. The toxins, TcdA and TcdB were studied intensively as major *C. difficile* virulence factors. *C. difficile* toxins A and B belong to the family of large clostridial cytotoxins. They are often called clostridial glucosylating toxins, as their toxic potency depends on their glucosyltransferase activity (Von Eichel-Streiber *et al.,* 1996; Schirmer & Aktories 2004; Aktories & Barbieri 2005; Aktories & Just 2005; Voth Ballard 2005). *C. difficile* disease is caused by TcdA and TcdB and is detected in the stools of patients by antibody-based and cytotoxicity assays.

These two toxins, along with the other members of the large clostridial toxin family of toxins, target the Ras super family of small GTPases for modification via glycosylation. This irreversible modification inactivates these small regulatory proteins and disrupts the signaling pathways in the cell. Along with enzymatic modification of targets, there are also several important steps in receptor-binding and cell entry necessary for intoxication.

In in vivo studies of toxins, it was found that toxin B was unable to cause disease in the absence of toxin A. In contrast, purified toxin A was capable of causing disease even in the absence of toxin B (Lyerly *et al.,* 1985).

The experimental evidence for receptor-binding by the C-terminus of TcdA comes from a series of studies showing the neutralizing capacity of monoclonal antibodies to this region of the protein (Frey & Wilkins 1992) and the ability of recombinant fragments from this region to protect against the toxin. TcdA-induced cytotoxicity is blocked by a fragment of C-terminus by the competitive inhibition of receptor-binding (Sauerborn *et al.,* 1997). C-terminal receptor-binding domain was shown to be necessary for both receptor-binding and effective endocytosis of the holotoxin (Frisch *et al.,* 2003).

The C-terminal Repetitive Domain (CRD) binds carbohydrates on colonic epithelial cells as an initial step in pathogenesis. Various oligosaccharides, including the linear B-type 2-trisaccharide a-Gal-(1, 3)-b-Gal-(1, 4)-b-GlcNAc, bind specifically to TcdA (Tucker & Wilkins 1991; Teneberg *et al.,* 1996). The functional importance of carbohydrates binding to TcdA is supported by previous studies. TcdA-f2 consists of nine Short Repeats (SR) and two Long Repeats (LR). First, diethylpyrocarbonate modification of histidine residues in TcdA specifically abolishes cytotoxicity and receptor-binding activities (Roberts & Shone 2001).

Ho et al. (2005) solved the crystal structure of a 127-aa fragment of C-terminal repeats from *C. difficile* toxin A with the 1.85 A° resolution. The structure of a 127-residue C-terminal fragment of TcdA (TcdA-f1) is highly repetitive which contains SR and LR. The crystal structure of TcdA-f1 reveals that each SR or LR contains a single β-hairpin consisting of a pair of five- to six-residue antiparallel β-strands connected by a tight turn. The first short repeat (SR1) starts from Gly14 residue and ends with Gly33 residue. The second short repeat (SR2) consists of Arg64 residue to Thr83 residue. The third short repeat (SR3) starts from Gly84 and ends at Gly105. The fourth short repeat (SR4) starts from Gly106 to Gly126. Each SR consists of two-beta strands connected by a turn and one repeat unit is connected with other repeat by a loop. There is an LR, which is in between the two short repeats SR1 and SR2. The LR starts from Gly33 residue and ends with Arg64 residue. It has also two strands connected by a turn.

The 3D structure of TcdA-fl suggests that the boundaries of each SR or LR should be defined to coincide with the beginning of the β-hairpin and the end of the connecting loop preceding the following β-hairpin. Each β-hairpin interacts with both the preceding and following β-hairpins, except for the N- and C-terminal hairpins. The N-terminal end of TcdA-fl adopts a non-natural structure due to the truncation of the protein. There is a hairpin in N-terminal hairpin, comprising hydrophobic residues which form a non-natural intermolecular interface with a few exposed hydrophobic residues of C-terminal.

In the present work, we have done molecular dynamics simulation and Principal Component Analysis (PCA) of C-terminal domain from *Clostridium difficile* toxin A to understand its global and local motional properties which will help to understand its function and mode of action.

6.2: Materials and Methods

The 1.85 A° resolution x-ray structure of *C. difficile* toxin A (Protein Data Bank code 2f6e) (Ho *et al.*, 2005) was used as a starting structure. A single monomer was solvated with SPC water molecules in a cubic box, having an edge length of 35 A°. The simulation was performed using GROningen MAchine for Chemical Simulation (GROMACS) (Lindahl *et al.*, 2001). The LINCS algorithm was used to constrain all bond lengths (Hess *et al.*, 1997). The simulation was conducted at a constant temperature (300 K), coupling each component separately to a temperature bath using the Berendsen coupling method (Berendsen *et al.*, 1984). A cutoff of 0.9 nm was used for Lennard Jones interaction and 1.0 nm for Coulomb interaction (Darden *et al.,* 1993). The time step was 2 fs, with coordinates stored after every 4 ps. MD simulation was performed for 3 ns. Before running simulation, an energy minimization was performed in steepest descent method (converged at 523 steps), followed by conjugate gradient method (converged at 8 steps) (Essmann *et al.,* 1995; Bothra *et al.,* 1998); and this was followed by 1.0 ns of simulation imposing positional restraints on the non-H atoms. The positional restraints were then released and 3 ns production run were obtained and analyzed. Analysis programs from GROMACS were used and PCA was performed with the MD trajectory.

6.3: Results

The overall structural stability of the protein during the simulation was monitored using several parameters like Root Mean Square Deviation (RMSD), Radius of gyration (Rg), Root Mean Square Fluctuation (RMSF), etc.

The time evolution of RMSD is computed taking the constrained structure of the whole protein as initial structure and presented in Fig.1.

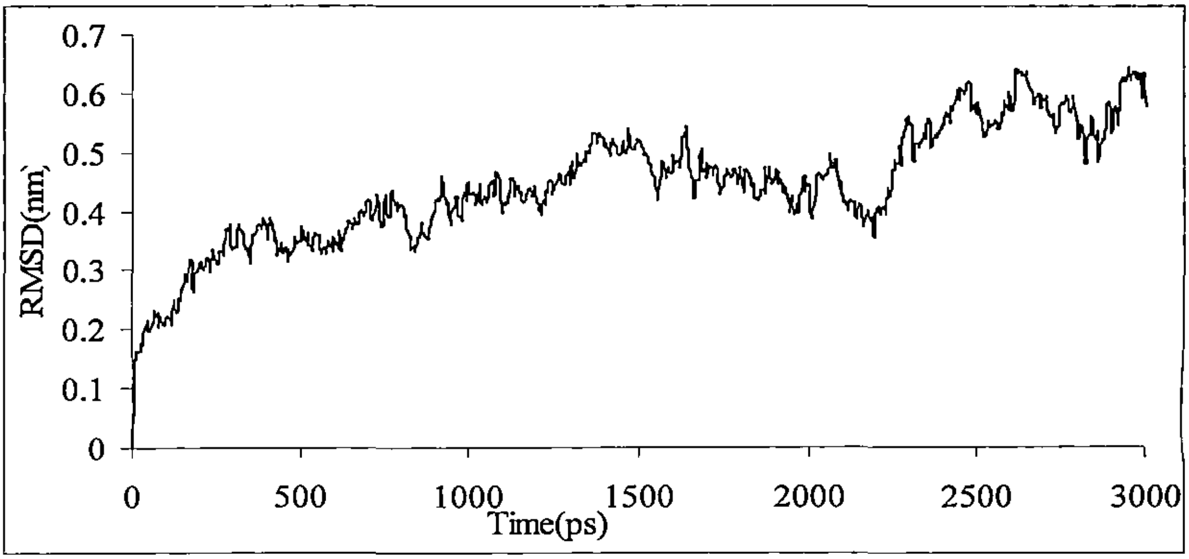


Figure-1: The time evolution of RMSD during 3000ps dynamics simulation time.

The time evolution of Rg is presented in Fig.2,

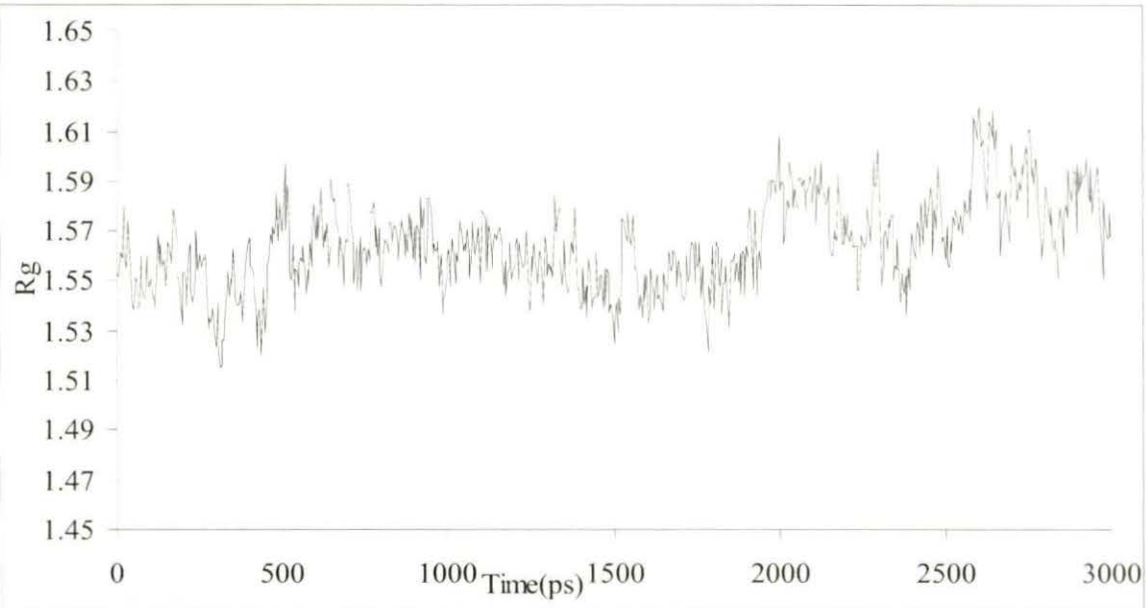


Figure-2: Time evolution radius of gyration changes in aqueous medium during 3000ps dynamics simulation.

RMSF indicates the flexibility of the protein. RMSF of Cα and B factors are presented as a function of residue numbers in Fig.3.

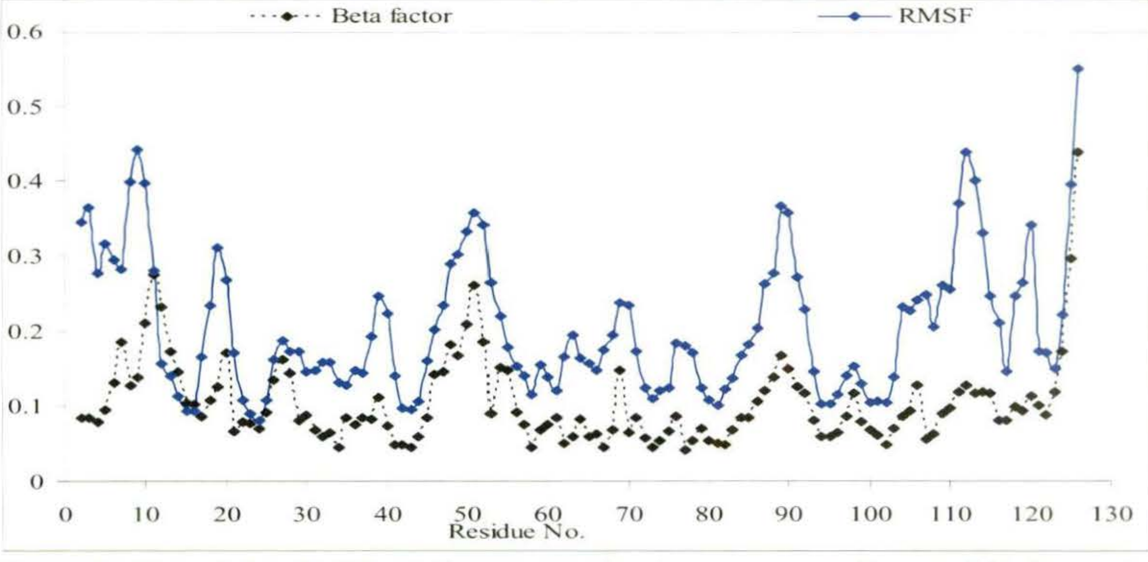


Figure-3: Plot of the RMSF of Cα atoms value in aqueous medium and B factors is presented as a function of residue number of *Clostridium difficile* toxin A in starting X-ray structure.

We have taken some selected snapshots from the dynamics trajectory considering time evolution of RMSD as a guideline. The snapshots were taken when the RMSD from initial structure was high, and they are represented in Fig.4.

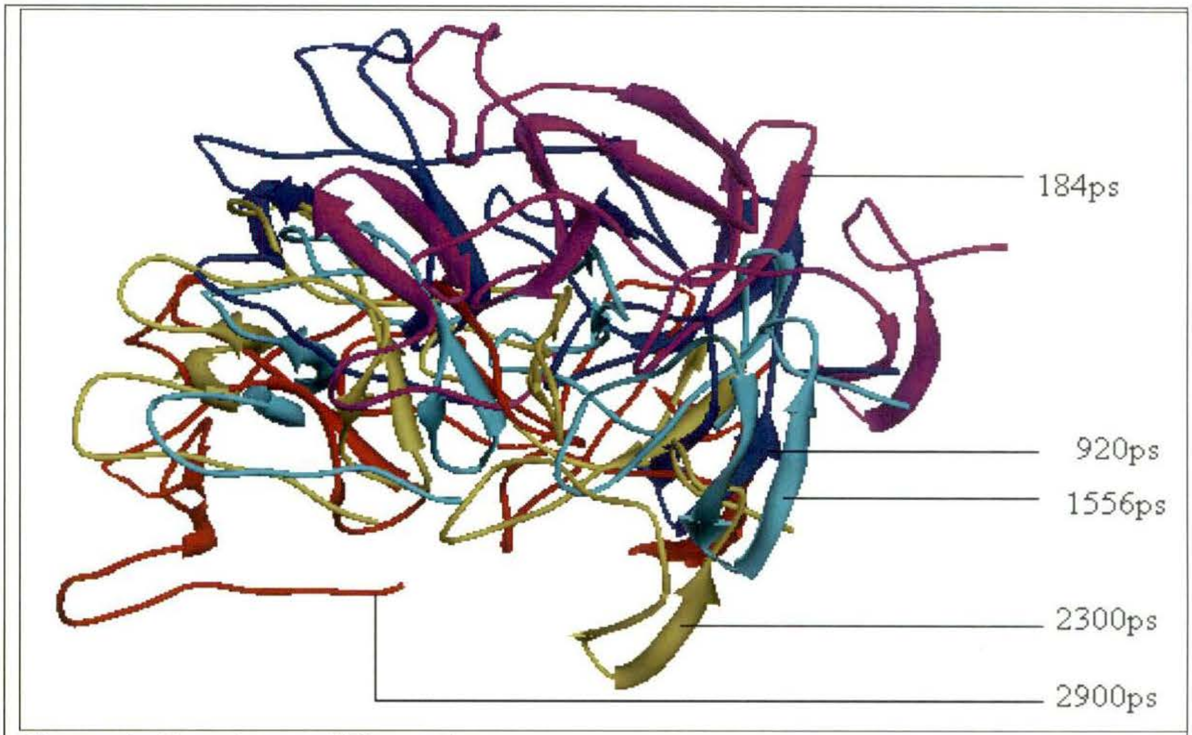


Figure-4: Snapshots at different time.

We have also calculated the RMSD of the SR1 and LR. We have determined the RMSD values of the SR and LR and their structural elements, which is given in Table 1.

Table 1: Mean RMSD values of different repeating unit including strand, turn and Loop

Repeat Unit	Overall	Strand I	Strand II	Turn	Loop
SR1	0.196788	0.107284	0.173618	0.145548	0.170353
LR	0.236153	0.103893	0.110976	0.115819	0.254105
SR2	0.239528	0.212295	0.139324	0.169818	0.216462
SR3	0.170861	0.124877	0.143989	0.123328	0.165811
SR4	0.230572	0.109931	0.141252	0.104894	0.245512

Time evolution of RMSD for different repeating units are presented in Fig.5.

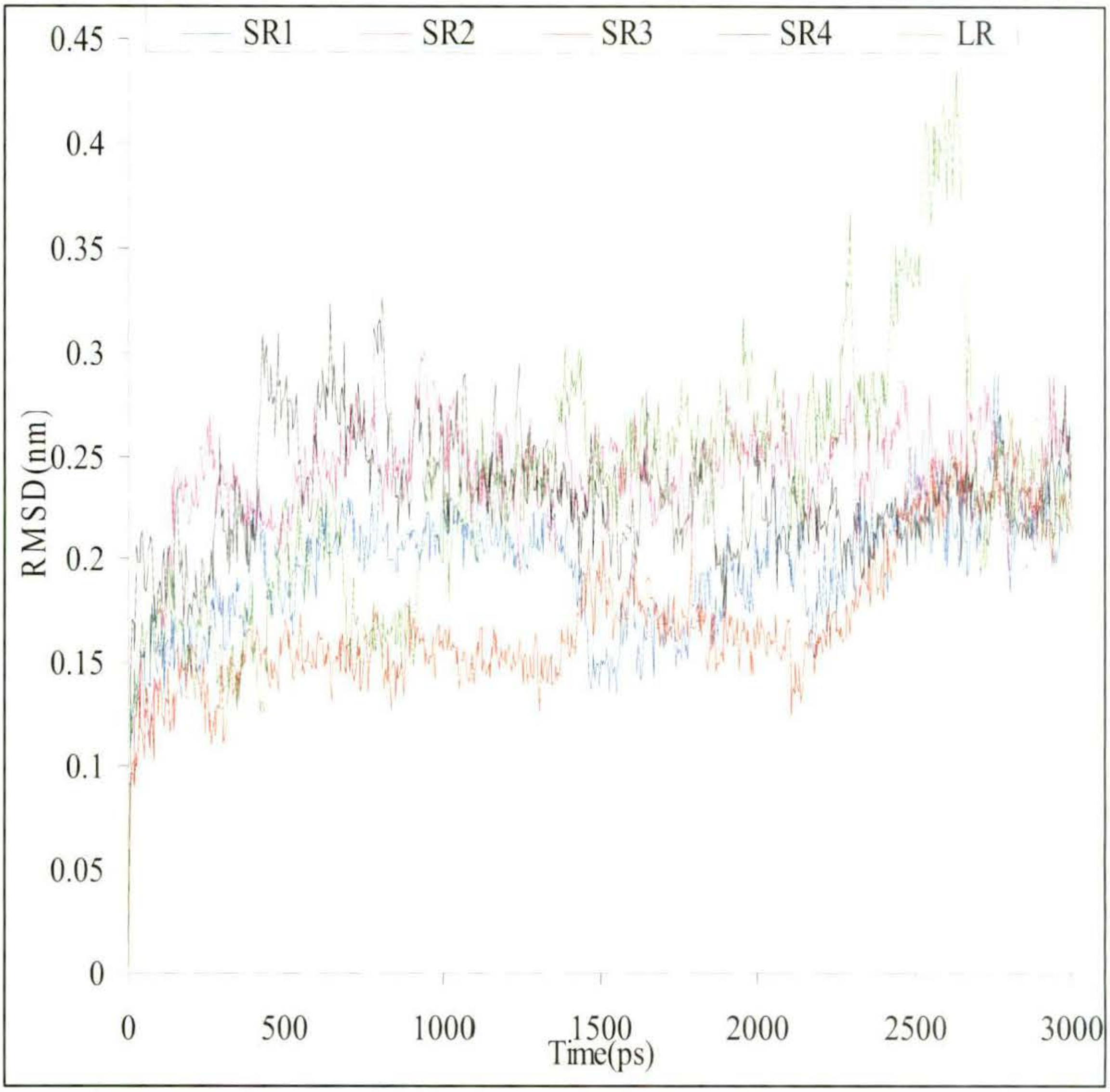


Figure-5: Time evolution of RMSD of different repeating unit of *Clostridium difficile* toxin A.

The average RMSD for the overall strands of different repeating units are represented in a histogram (Fig.6) and it is seen that strand I of SR2 and strand I of LR showed the highest and lowest fluctuations respectively among all strands. Time evolution of RMSD values of strand I of SR2 and strand I of LR are given in Fig.7.

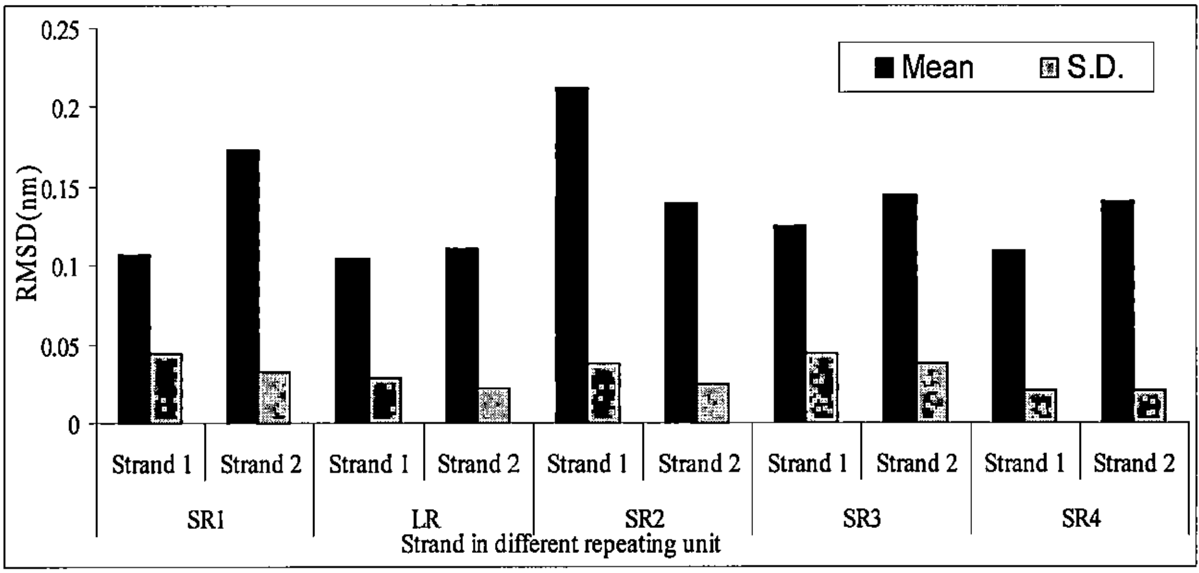


Figure-6: Histogram of RMSD for the all the strands of different repeating unit

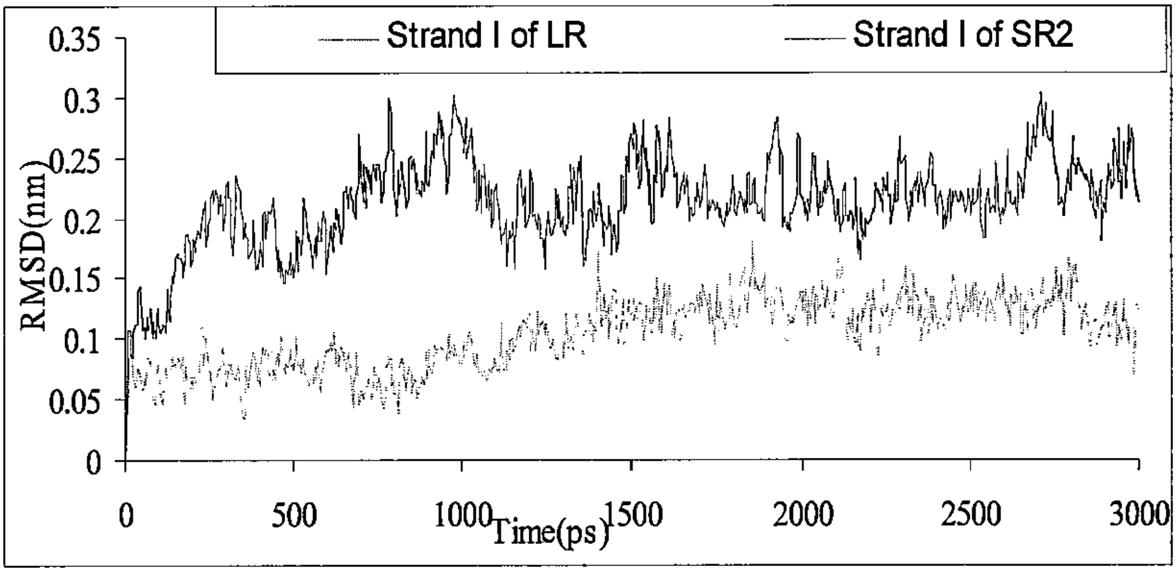


Figure-7: Plot of time evolution of RMSD values of SR2 of strand I and LR of strand I.

A common approach in the identification of the major motions of a protein is the use of PCA (Garcia, 1992; and Amadei et al., 1993). PCA reduces the dimensionality of a complex data set and applied to decompose a complex motion of proteins, which are characterized by an eigenvector and an eigenvalue. The eigenvalue for a given motion represents the contribution of the corresponding eigenvector to the global motion of the protein. PCA of the C. difficile toxin A simulation reveals that the first 10 eigenvectors account for 87.16%

of the global motion and that the first eigenvector corresponds to 49.68% of the total motion, the second to 15.32%, and the third to a further 5.90% (Fig.8).

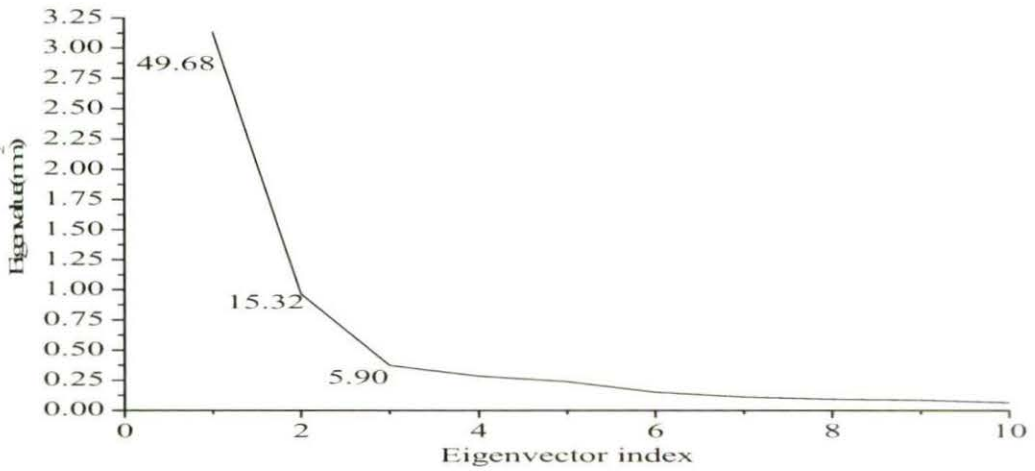


Figure-8: Plot of eigenvalues with eigenvector indexes.

The protein comprises one LR and four SR. Each repeat consists of two beta strands followed by a loop.

The projection of RMSF on four vectors is presented in Fig.9.

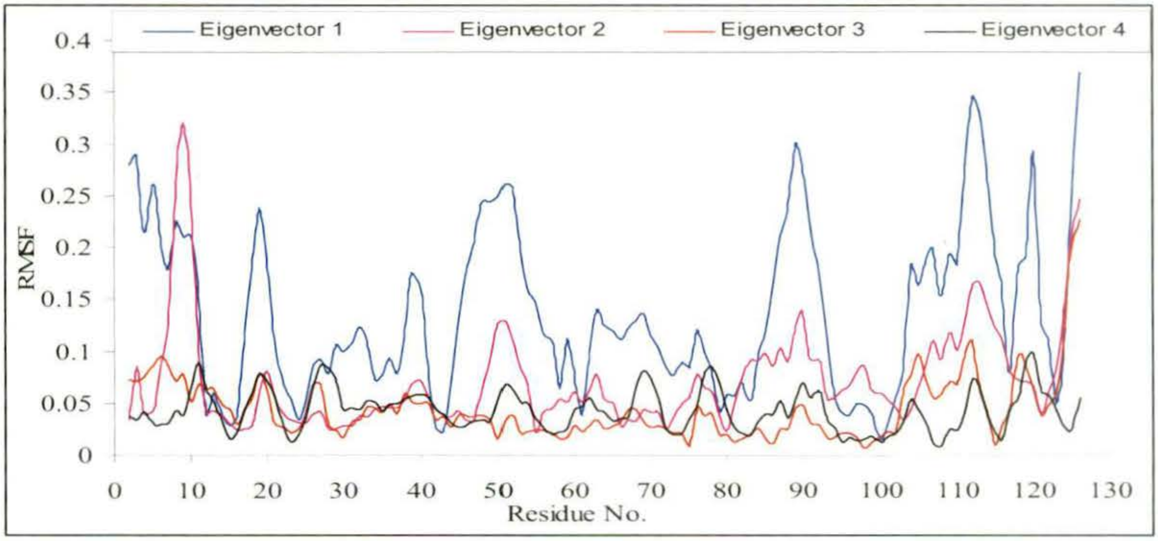


Figure-9: Plot of the projection of RMSF on four vectors.

Time evolution of principal component 1 (PC1), principal component 2 (PC2), principal component 3 (PC3) and principal component 4 (PC4) in water is represented in Fig.10.

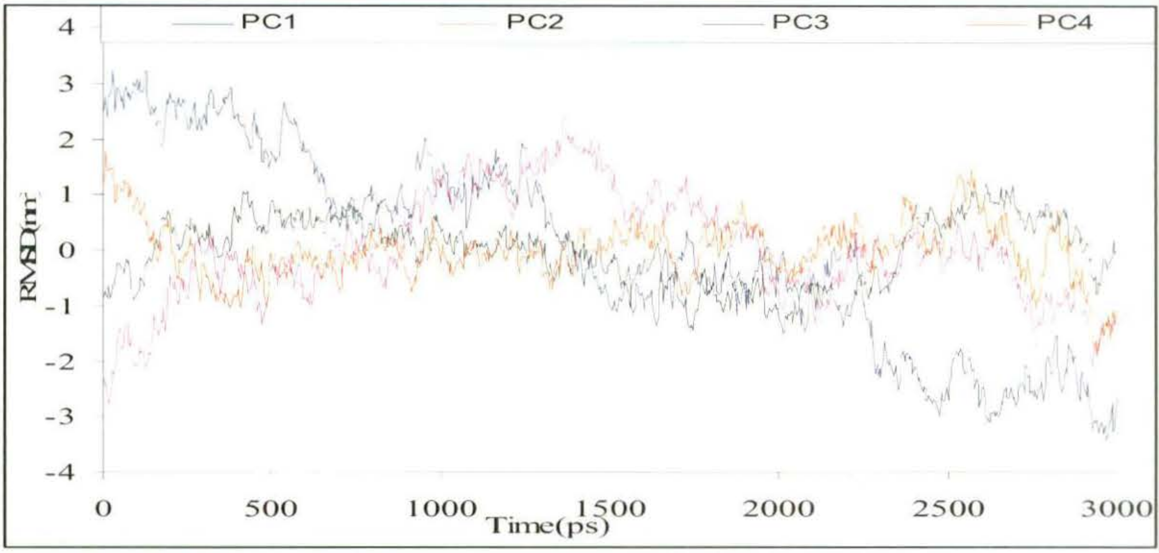


Figure-10: Plot of four principle components (nm) with simulation time

The projection of the dynamics trajectory onto the first two eigenvectors was analyzed and presented in Fig.11.

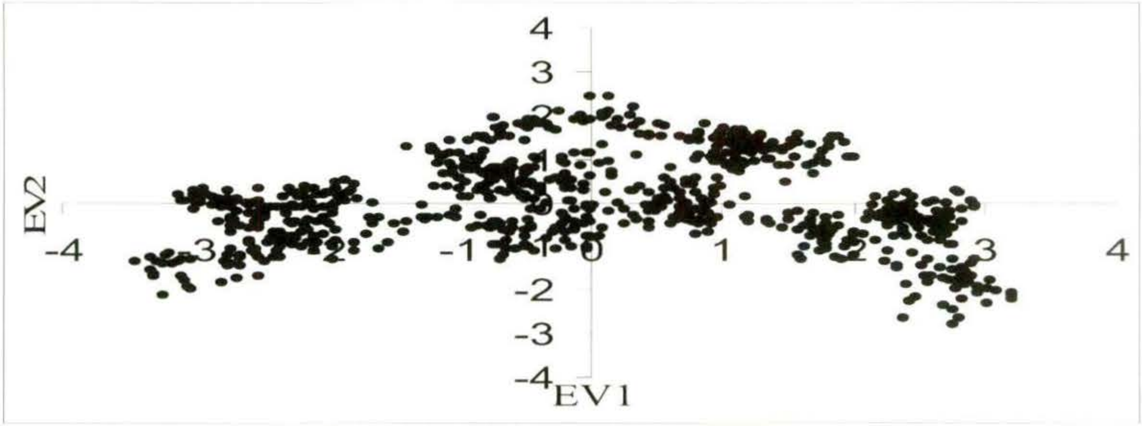


Figure-11: Plot of eigenvector1 with eigenvector 2.

According to the report of Greco et al. (2006), the crystal structure of the C-terminal 255 residues of TcdA from C. difficile strain 48489, toxinotype VI (TcdA-f2) found to be a synthetic derivative of the natural carbohydrate receptor, a-Gal-(1, 3)-b-Gal-(1, 4)-b GlcNAcO(CH2)8CO2CH3 (CD-grease). The structural feature and sequence similarity of LR as well as the similar binding mode indicates the conserved nature of the seven carbohydratebinding sites in TcdA-f2 which consist of an LR and the hairpin turn of the following SR.

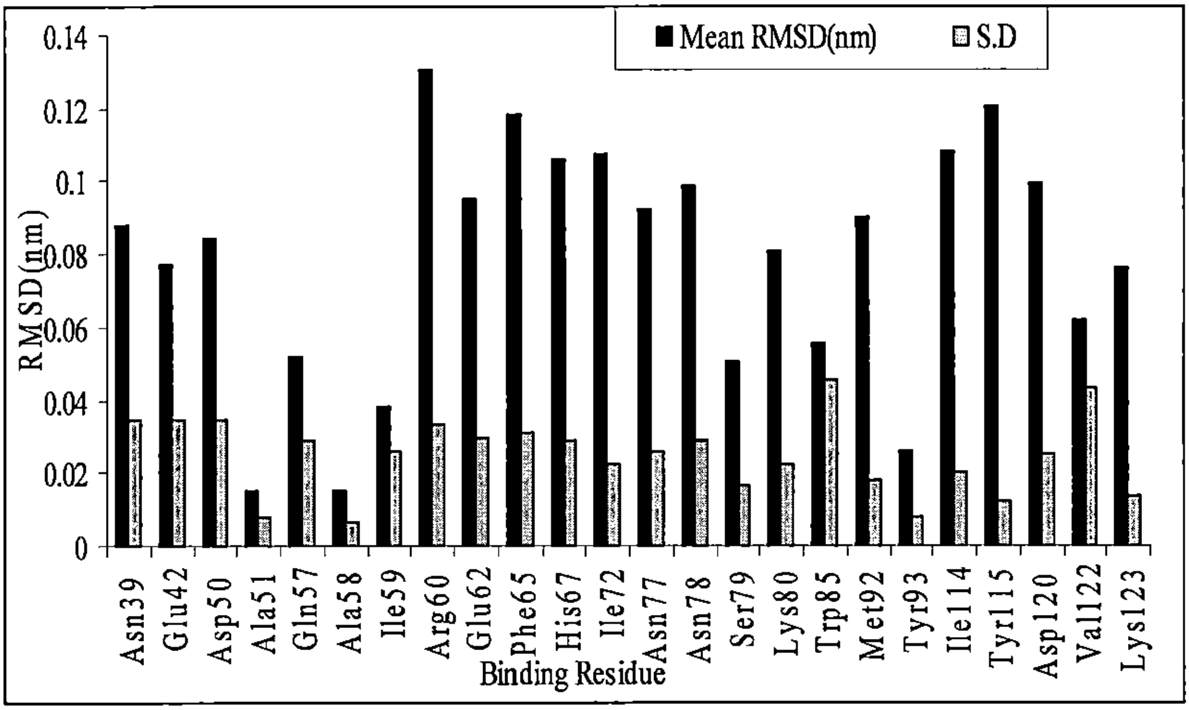


Figure-12: Histogram of RMSD values for the residues of binding site during the simulation time

To get insight about the binding mechanism, the residues of binding site were critically analyzed. The residues Lys80, Arg60, Ser79, Asp50, and Gln57 are very important for binding processes, as pointed out by Greco et al. (2006). From the RMSD values during the simulation time (Fig.12), it is evident that the residues, Ala51, Ala 58, Ile59 and Tyr93 have very low fluctuation. The residues, Arg60, Phe65, His67, Ile72, Ile114 and Tyr115 have high fluctuation, and the fluctuation was moderate for the residues Asn39, Glu42, Asp50, Gln57, Ser79, Trp85, Val122 and Lys123 during the whole simulation time (Mean RMSD and SD value is given in Table 2).

Table-2: Mean RMSD and S.D. of different residue binding site.

Binding residue	Mean RMSD(nm)	S.D
Asn39	0.087781	0.034847
Glu42	0.077564	0.034463
Asp50	0.084088	0.034517
Ala51	0.01535	0.007662
Gln57	0.052091	0.028836
Ala58	0.015217	0.006695
Ile59	0.038305	0.026329
Arg60	0.130462	0.033052
Glu62	0.095121	0.029304
Phe65	0.118562	0.031224
His67	0.106327	0.028664
Ile72	0.107694	0.022652
Asn77	0.09213	0.025975
Asn78	0.098776	0.02911
Ser79	0.050197	0.016265
Lys80	0.080878	0.022682
Trp85	0.055926	0.045661
Met92	0.090154	0.018117
Tyr93	0.026039	0.007814
Ile114	0.108123	0.020269
Tyr115	0.120166	0.012418
Asp120	0.099466	0.02529
Val122	0.062328	0.043586
Lys123	0.076151	0.01389

The 6-OH of β-galactose accepts hydrogen from Lys80 which has RMSD value 0.080878. The B-face of β-galactose packs against the apolar proximal portions of Arg60 which has RMSD value 0.130462, and Ser79 has RMSD value of 0.050197. The α -galactose at the nonreducing end lies in a pocket formed by highly conserved residues in the loop after the β-hairpin of the LR. Its hydrophobic B-face packs against Ile59 with an average RMSD value 0.038305 and the 6-OH forms hydrogen bonds with Asp50 (RMSD 0.084088) and Arg60 (RMSD 0.130462), whereas the 4-OH accepts hydrogen from Gln57 (RMSD 0.052091). During the simulation, several hydrogen bonds broke and formed. It is found that the number of hydrogen bonds ranged from 54 to 85 (Fig.13).

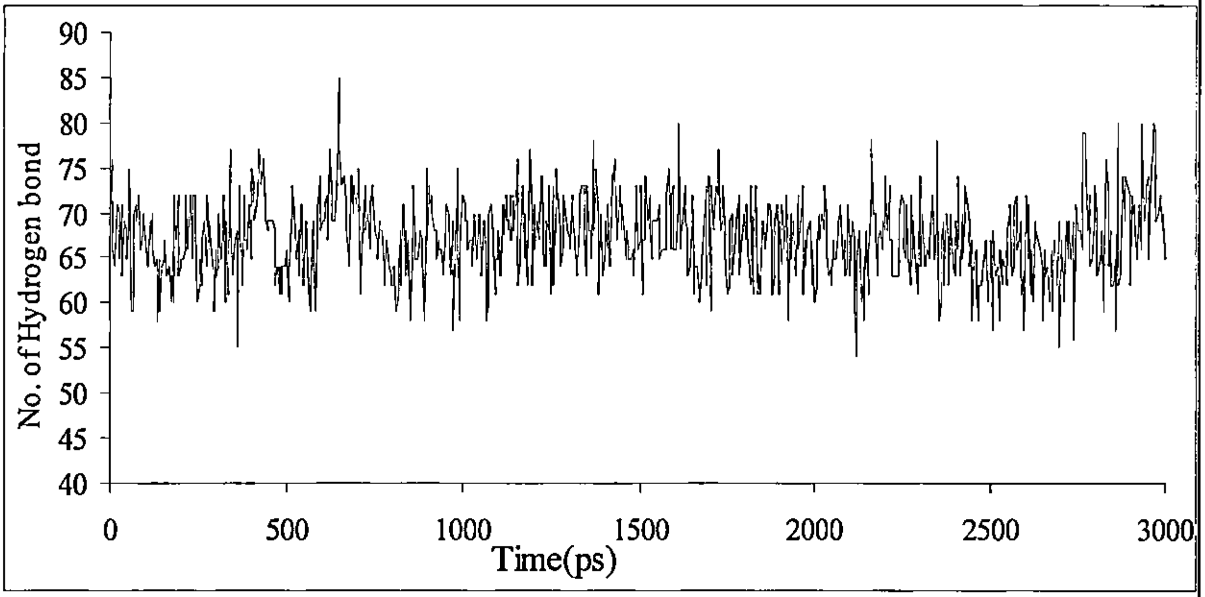


Figure-13: Number of hydrogen bond during the whole simulation time.

A series of hydrogen bonds were observed in the side chains of highly conserved amino acids of 18 residue loop of LR in C. difficile toxin A (Ho *et al.*, 2005). We have monitored the distance of several hydrogen bonds during the simulation time. Some of the hydrogen bonds were very strong, but some were broken during the simulation time. It was found that the hydrogen bonds between amide nitrogen of Asn53 and the carbonyl oxygen atoms of Pro46 (H1), amide nitrogen of Asn53 and the carbonyl oxygen atoms of Ile 54 (H2), carbonyl oxygen of Asn53 and the amide nitrogen atoms of Thr49 (H3), carbonyl oxygen of Asn53 and the amide nitrogen atoms of Asp50 (H4) and the side chain of Thr49 and the main chain peptide groups of Pro46 (H5) remain intact during the whole simulation time, whereas the hydrogen bonds between side chain of Glu55 and the main chain peptide groups of Ala47 (H6), the side chains of Gln57 and the main chain peptide groups of Ile 54 (H7 and H8) and hydrogen bonds formed between the side chains of Asn48 and Asn52 (H9), as well as between the side chains of Asp50 and Arg60 (H10 and H11), were broken during the whole simulation time (Fig.14).

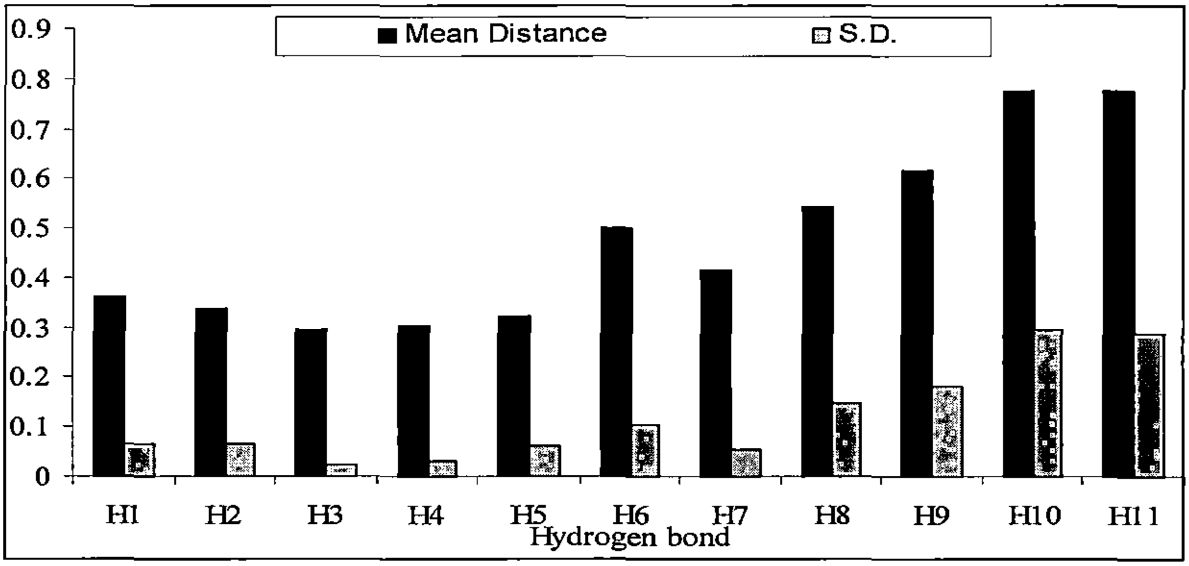


Figure-14: Plot of different mean hydrogen bond distance in the whole simulation time

4. Discussion

It is evident from Fig.1 that RMSD increased slowly up to 2,000 ps, then there was a small jump at about 2,500 ps and remained at that state with low fluctuation. The initial drift in RMSD may be due to the difference of crystal structure with solution structure. From Fig.2, it is clear that Rg does not show much variation during the simulation time, which indicates that the protein is not much flexible.

The plot of B-factor and RMSF (Fig.3) presents a similar trend across the sequence, although there are differences in the environmental condition of the protein molecule in the x-ray and MD studies. From RMSF, it is evident that the first and last residue fluctuates considerably. Interestingly, pronounced fluctuations are observed along some amino acid stretches (16-23, 45-57, 84-94, and 110-117), which indicate the flexibility of the toxin in that region.

From the snapshots (Fig.4), it is also clear that there was no major change in the protein conformation.

Among the SR units, SR1 and SR3 have low RMSD values. Strand I and turn of SR2 have approximately double RMSD in comparison to the strand I of other SR. Strand II of SR1 has higher RMSD value amongst the strand II of the entire SR. The loop that connects LR and SR2 has the highest RMSD value among different loops (see Table-1).

It is clear from the Fig.5 that SR1 and SR3 have less fluctuation among the SR.

From Fig.9, it is evident that fluctuation is high in the projection on 1st vector. It indicates the analysis of 1st vector will provide more information regarding the collective motion of the protein.

Fig.10 indicates PC1 fluctuates remarkably in comparison to PC2, PC3 and PC4.

From Fig.11, it is clear that the toxin is sampling different conformational space during the simulation. It is also clear from the Fig.11 that the protein traverses one conformational space around the origin, second one at the right side of the origin, and the third one at the left side of the origin which are not much scattered, indicating less conformational freedom of the toxin; this is also revealed by different snapshots of the different structures extracted along the simulation trajectory in different times (Fig.4).

The hydrogen bond network is almost conserved during the simulation time. Extensive study of the residues of binding site was performed in order to understand the binding mechanism. From Table 2, it was found that Ser79 and Gln57 have low RMSD values. The Lys80 residue, which accepts hydrogen from the 6-OH of β-Galactose and donates hydrogen to a water molecule, has a moderate RMSD value indicating the conformation change of this residue is possible during binding process. Asp50 has a moderate RMSD value. Arg60 has much higher flexibility than the other binding residues (Lys80, Ser79, Asp50, and Gln57). The side chain of Gln57 is involved in hydrogen bond with main chain peptide group of Ile54. Arg60 has hydrogen bond with Asp50 which often breaks during simulation and has high conformational freedom, and this is also supported by its RMSD value.

The binding site is much flexible so the binding may be initiated by those residues which have a very crucial role in inducing binding ligand. The other residues in binding site may be responsible for the stability of ligand toxin complex. These analyses will further help to understand the binding mechanism and overall motional properties of the protein. From the overall study of the dynamics trajectory such as Rg, RMSD of the toxin and PCA, it is clear that *C. difficile* toxin A is not much flexible.

6.5: References

Aktories K & Barbieri JT (2005) Bacterial Cytotoxins: Targeting Eukaryotic Switche. *Nat. Rev. Microbiol.,* **3**: 397-410.

Aktories K & Just I (2005) Clostridial Rho-Inhibiting Protein Toxins. *Curr. Top. Microbiol. Immunol.,* **291**:113-145.

Amadei A, Linssen ABM & Berendsen HJC (1993) Essential Dynamics of Proteins. *Proteins,* **17**:412-425.

Aronsson B, Mollby R & Nord CE (1981) Occurrence of Toxin Producing *Clostridium difficile* in Antibiotic-Associated Diarrhea in Sweden. *Med. Microbiol. Immunol.*(Berlin), **170**:27-35.

Bartlett JG, Moon N, Chang TW, Taylor N & Onderdonk AB (1978) Role of *Clostridium difficile* in Antibiotic-Associated Pseudomembranous Colitis. *Gastroenterology.* **75**:778-782.

Bartlett JG, Taylor NS, Chang T & Dzink J (1980a) Clinical and Laboratory Observations in *Clostridium difficile* Colitis, *Am. J. Clin. Nutr.,* **33**:2521-2526.

Bartlett JG, Tedesco FJ, Shull S, Lowe B & Chang T (1980b) Symptomatic Relapse After Oral Vancomycin Therapy of Antibiotic-Associated Pseudomembranous Colitis. *Gastroenterology*, **78**:431-434.

Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A & Haak JR (1984) Molecular Dynamics With Coupling to an External Bath. *J. Chem. Phys.,* **81**:3684-3690.

Bothra AK, Roy S, Bhattacharyya B & Mukhopadhyay C (1998) Molecular Dynamics Simulation of Colchicinoids. *Journal of Bimolecular Structure and Dynamics,* **15**:999-1008.

Burdon DW, George RH, Mogg GA, Arabi Y, Thompson H, Johnson M, Alexander-william J & Keighley MR (1981) Faecal Toxin and Severity of Antibiotic-Associated Pseudomembranous Colitis. *J. Clin. Pathol.,* **34**: 548-551.

Darden T, York D & Pedersen L (1993) Particle Mesh Ewald—An N.log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.,* **98**:10089-10092.

Don GJ & Devis AE (1981) The Association Between Antibiotic-Associated Diarrhoea and *C. difficile* Toxin in Children. *Aust. N. Z. J. Med.,***11**:433-434.

Essmann U, Perera L, Berkowitz ML, Darden T, Lee H & Pedersen LG (1995) A smooth particle mesh Ewald method. *J. Chem. Phys.,* **103**:8577–8593

Frey S M & Wilkins T D (1992) Localization of Two Epitopes Recognized by Monoclonal Antibody PCG-4 on *Clostridium difficile* Toxin A. *Infect. Immun.,* **60**:2488-2492.

Frisch C, Gerhard R, Aktories K, Hofmann F & Just I (2003) The Complete Receptor-Binding Domain of *Clostridium difficile* Toxin A is Required for Endocytosis. *Biochem. Biophys. Res. Commun.,* **300**:706-711.

Garcia AE (1992) Large-Amplitude Nonlinear Motions in Proteins. *Phys. Rev. Lett.,* **68**:2696-2699.

George WL, Sutter VL, Goldstein EJ, Ludwig SL & Finegold SM (1978) Aetiology of Antimicrobial- Agent-Associated Colitis. *Lancet,* **311**:802-803.

George WL, Rolfe RD, Harding GK, Klein R, Putnam CW & Finegold SM (1982) *Clostridium difficile* and Cytotoxin in Feces of Patients with Antimicrobial Agent-Associated Pseudomembranous Colitis. *Infection,* **10**:205-208.

Greco A, Ho JGS, Lin S, Palcic MM, Rupnik M & Ng KK (2006) Carbohydrate Recognition by *Clostridium difficile* Toxin A. *Nature Structural & Molecular Biology,* **13**:460-461.

Hess B, Bekker H, Berendsen HJC & Fraaije JGE (1997) LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**:1463-1472.

Ho JGS, Greco A, Rupnik M and Ng KK-S (2005) Crystal Structure of Receptor- Binding C-Terminal Repeats from *Clostridium difficile* Toxin A. *PNAS,* **102**:18373-18378.

Kelly CP & LaMont JT (1998) *Clostridium difficile* Infection. *Annu. Rev. Med.,* **49**:375-390.

Lindahl E, Hess B & van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Modeling* **7**:306-317.

Lyerly DK, Saum KE, MacDonald DK & Wilkins TD (1985) Effects of *Clostridium difficile* Toxins Given Intragastrically to Animals. *Infect. Immun.,* **47**:349-352.

Meyers S, Meyers L, Bottone E, Desmond E & Janowitz HD (1981) Occurrence of *Clostridium difficile* Toxin During the Course of Inflammatory Bowel Disease. *Gastroenterology,* **80**:697-670.

Roberts AK & Shone CC (2001) Modification of Surface Histidine Residues Abolishes the Cytotoxic Activity of *Clostridium difficile* Toxin A. *Toxicon,* **39**:325-333.

Sauerborn M, Leukel P & von Eichel-Streiber C (1997) The C-Terminal Ligand- Binding Domain of *Clostridium difficile* Toxin A (TcdA) Abrogates TcdA-Specific Binding to Cells and Prevents Mouse Lethality. *FEMS Microbiol. Lett.,* **155**:45-54.

Schirmer J & Aktories K (2004) Large Clostridial Cytotoxins: Cellular Biology of Rho/Ras-Glucosylating Toxins. *Biochim. Biophys. Acta.,* **1673**:66-74.

Tedesco FJ (1981) Antibiotic-Associated Colitis: An Abating Enigma. *J. Clin. Gastroenterol.,* **3**:221-224.

Teneberg S, Lönnroth I, Torres López JF, Galili U, Halvarsson MO, Angstrom J & Karlsson K (1996) Molecular Mimicry in the Recognition of Glycosphingolipids by Gal3Galß4GlcNAcß-Binding *Clostridium difficile* Toxin A, Human Natural Anti-

Galactosyl IgG and the Monoclonal Antibody Gal-13: Characterization of a Binding-Active Human Glycosphingolipid, Non-Identical with the Animal Receptor. *Glycobiology*, **6**: 599-609.

Tucker KD & Wilkins TD (1991) Toxin A of *Clostridium difficile* Binds to the Human Carbohydrate Antigens I, X, and Y. *Infect. Immun.,* **59**:73-78.

Von Eichel-Streiber C, Boquet P, Sauerborn M and Thelestam M (1996) Large Clostridial Cytotoxins—A Family of Glycosyl Transferases Modifying Small GTPBinding Proteins. *Trends Microbiol.,* **4**:375-382.

Voth D E & Ballard JD (2005) *Clostridium difficile* Toxins: Mechanism of Action and Role in Disease. *Clin. Microbiol. Rev.,* **18**:247-263.

Willey SH & Bartlett JG (1979) Cultures for *Clostridium difficile* in Stools Containing a Cytotoxin Neutralized by Clostridium sordellii Antitoxin. *J. Clin. Microbiol.,* **10**: 880-884.

Chapter VII

Similarities and differences in motional properties of Staphylococcus aureus exfoliative toxins A and B as revealed by their Molecular dynamics simulation

7.1: Introduction

Staphylococcus aureus can cause a spectrum of exfoliative skin conditions including staphylococcal scalded skin syndrome (SSSS), which can present as a severe and life threatening illness in extremely premature neonates. Kapoor *et al* (2008) describe a case of an extremely premature neonate with SSSS and discuss relevant pathology (Kapoor *et al.*, 2008). The term Staphylococcal scalded skin syndrome (SSSS) is used for a collection of blistering skin diseases induced by the exfoliative (epidermolytic) toxins (ETs) of *Staphylococcus aureus* (Lowney *et al.*, 1967). It primarily affects neonates and young children although adults with underlying diseases are also susceptible. Staphylococcal infections are increases in all age groups worldwide, and show an increasing resistance to conventional antibiotics; despite the availability of a wide range of antibiotics, these infections still carry a significant morbidity and mortality, particularly among adults (Jeyakumari *et al.*, 2009). Because of the relative rarity of the disease and ease of treatment of SSSS has not received as much attention as it deserves by either clinicians or researchers. Furthermore, although the condition was described over a century ago, understanding of it began only when the toxins were discovered (Scopes *et al.*, 1974; Melish & Glasgow 1970; Kapral & Miller 1971). Even now, their mechanism of action is still not certain. However, recent data from computer modeling and crystallography of the toxins has provided us with a clearer and more defined approach to understanding the pathologic processes of the disease (Ladhani *et al.*, 1999). Two serologically distinct forms of exfoliative toxin exist and have been designated as ETA and ETB (Arbuthnott & Billeliffe, 1976; Kondo *et al.*, 1973). The two forms differ in molecular weight, isoelectric point (pI), and amino acid composition (Arbuthnott *et al.*, 1976; Johnson-Winegar & Spero 1983; Wiley & Rogolsky 1977).

X-ray crystallographic structures of ETA and ETB (Cavarelli *et al.,* 1997; Papageorgiou *et al.,* 2000; Vath *et al.,* 1997) suggest that the toxins are members of the trypsin-like serine protease family. Protease activity has not been demonstrated for either toxin in vitro, but both ETA and ETB have intrinsic esterase activity, which is associated with serine proteases (Baile & Redpath 1992). It is likely that both toxins are proteases. In addition to having possible protease activity, both ETA and ETB are reported to be bacterial super antigens (Vath *et al.,* 1997; Marrack & Kappler 1990; Monday *et al.,* 1999; Morlock *et al.,* 1980). The protein ETA contains two domains (I and II) of similar structure, which are built around a six-stranded antiparallel b sheet folded into a β barrel. This architecture is well-known as the serine protease chymotrypsin-like fold, which has been illustrated by several high resolution structures (Perona & Craik 1995; Lesk & Fordham 1996).

ETB consists of two similar domains packed together to form a compact molecule of approximate dimensions 60 × 56 × 36 Å. The core of the molecule contains the "trypsin-like" serine protease fold. Domain 1 is composed of a β-barrel (comprising strands β2–β) surrounded by two short helices (α2 and α3) the C-terminal portion of the molecule that contains two short β-strands (β14 and β15) and a short helix (α5). Domain 2 contains a short amphipathic N-terminal helix (α1) on one side of a six-stranded twisted β-barrel structure (β8–β13) (Papageorgiou *et al.,* 2000). Both ETA and ETB share amino acid identity with staphylococcal V8 protease. This identity includes residues of the V8 protease Ser-His-Asp catalytic triad, a "signature sequence" common to serine proteases. Although this suggests that ETA and ETB are serine proteases, no protease activity has been demonstrated for either toxin to date. However, both ETA and ETB exhibit esterase activities, an activity commonly associated with proteases (Baileyet *et al.,* 1992). In trypsin-like serine proteases, the cleavage of the scissile bond involves formation of a negative charge that is stabilized in a pocket called the "oxyanion hole," formed by the main-chain atoms (involving the NH groups of Gly193 and Ser195) Surrounding the catalytic serine.

The oxyanion holes of both ETA and ETB are occupied, indicating the crystal structures represent an inactive form of the enzymes. The differing biological activities of the ETs, particularly the lethal and pyrogenic properties of ETB not shared by ETA, likely involve features that are not common between the toxins. The most apparent difference between the ETs is a 14-residue loop extension located before the C-terminal helix in ETB. This large loop is involved in a molecular dimer interface, although the biological relevance of this finding will need to be addressed further. Another potential feature that may be involved in the differing biological properties of the ETs is the small ligand-binding site found in ETA but not ETB. Given the differing structural features on the surface of the ETs, the binding of separate receptors or the same receptor with differing binding modes likely explains the dissimilar functions of the ETs.

Site directed mutagenesis shows that Tyr 157 and Tyr 159 are essential for toxicity. Mutation of any one Tyr reduces the toxicity while double mutation appeared to complete loss of exfoliative activity. Both these residues interact with S1 subsite of Glu182 and Val183. Although both toxins possess similar biological activity, the clustering of tyrosine residue is seen in rETB not in rETA. From the structure based sequence alignment of rETA and rETB in case of rETA tyrosine is not present in aligned position (Papageorgiou *et al.*, 2000)

Elucidating the mechanism of action of the toxins holds exciting prospects for understanding the normal physiology of the skin, targeting drugs to very specific regions of the skin, and developing antitoxins and toxoids that may soon play a vital role in the treatment and prevention of SSSS. (Ladhani *et al.,*1999). In the present work we have done molecular dynamics simulation and principal component analysis of *Staphylococcus aureus* exfoliative toxins A and B to understand their global and local motional properties which will help to understand it function, mode of action and differences between the toxin A and B.

7.2: Materials and methods

The 2.0A°and 2.8 A° resolution x-ray structure of *Staphylococcus aureus* exfoliative toxins A and B (Protein Data Bank code 1DUA and 1DT2 respectively) (Papageorgiou *et al.,* 2000) were used as starting structures. Each monomer was solvated with SPC water molecules in a cubic box having edge length of 35Å. The simulation was performed using GROMACS (Lindahl *et al.,* 2001). The LINCS algorithm was used to constrain all bond lengths (Hess *et al.,* 1997). A cutoff of 0.9 nm for Lennard-Jones interactions was used and the particle mesh Ewald method (Darden *et al.* 1993; Essmann *et al.,* 1995) was employed to calculate longer-range electrostatic contributions on a grid with 0.12-nm spacing and a cutoff of 1.0 nm. The simulation was conducted at constant temperature (300 K), coupling each component separately to a temperature bath using the Berendsen coupling method (Berenson *et al.,* 1984). The time step was 2 fs, with coordinates stored after every 4 ps. MD simulation was performed for four nano seconds for both the protein 1DUA & 1DT2. Before running simulation, an energy minimization was also performed for both the protein 1DUA & 1DT2 in steepest descent method (converged at 648 and 1050 steps respectively) and this was followed by 1.0ns of simulation imposing positional restraints on the non-H atoms. The positional restraints were then released and 4 ns production run were obtained and analyzed. Analysis programs from GROMACS were used and principal component analysis (PCA) was performed with the MD trajectory.

7.3: Results

The overall structural stability of the protein during the simulation has been monitored using several parameters like RMSD, radius of gyration (Rg), RMSF etc. The time evolutions of RMSD of the whole protein of Staphylococcus aureus exfoliative toxins A and B during the full simulation time (Fig.1) shows no major difference in case of the structural variation for toxin A and toxin B as measured by RMSD values. RMSD slightly increases upto 3000ps for toxin A and in case of toxin B, simulations become stable after 1000ps.

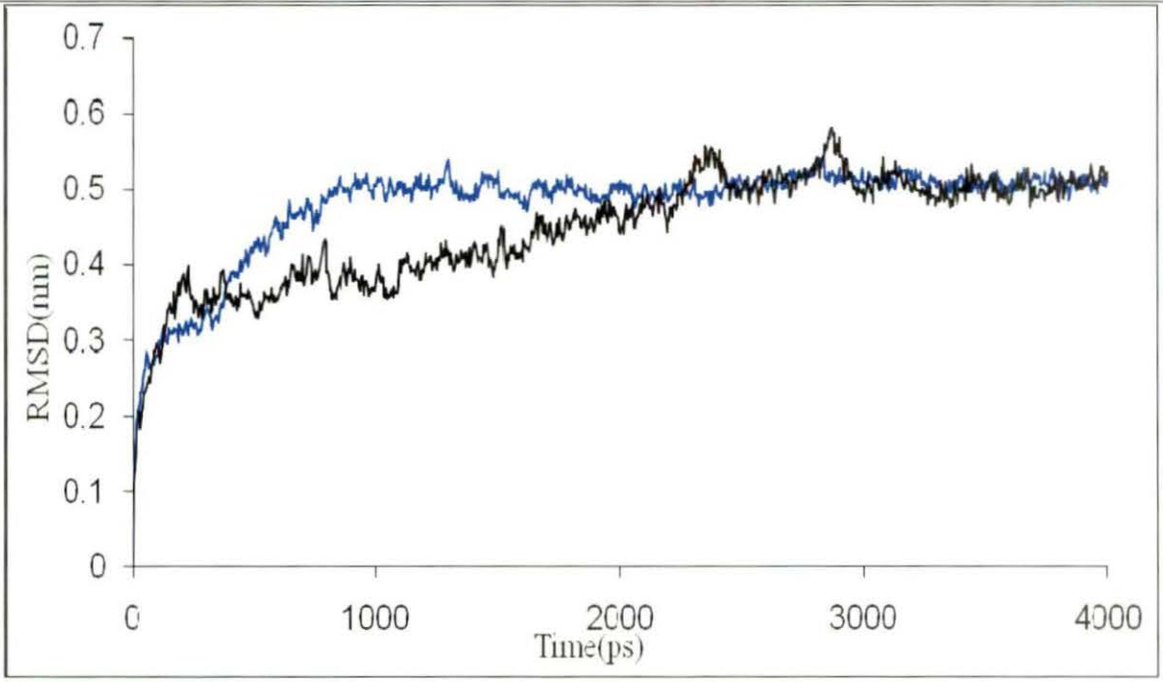


Figure-1: **RMSD:** Root mean square deviations(RMSD) of the protein as a function of time with respect to starting structure during the MD simulations are shown for toxin A i.e., 1DUA(Black) and toxin B i.e.1DT2(blue).

The variation of radius of gyration (Rg) as function of time (Fig. 2) indicates the stability of the simulation i.e. both toxins A and B do not change shape and size during the simulation time. However it is seen that Rg of toxin A remains greater than toxin B during the whole simulation time.

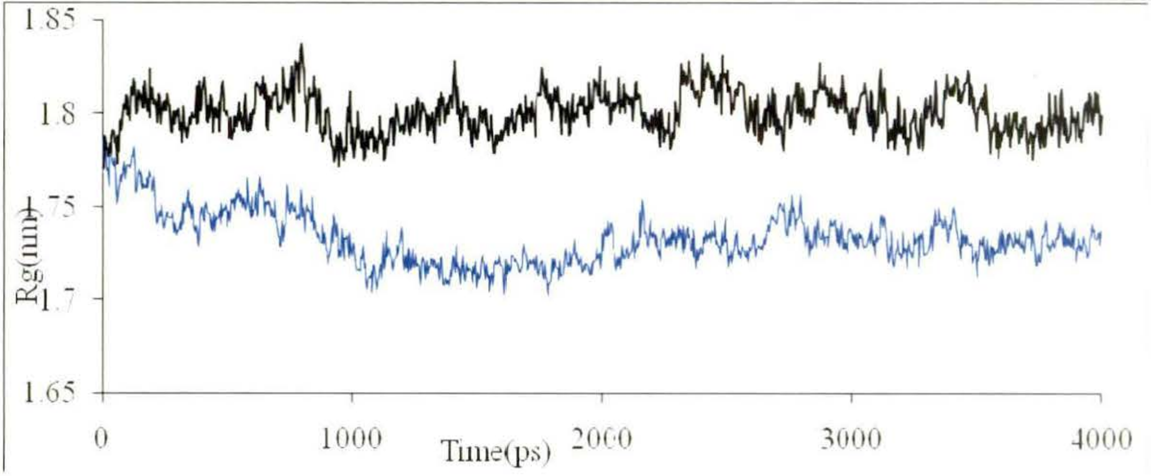


Figure-2: **Radius of gyration**: Radius of gyration (Rg) as a function of time with respect to starting structure during the MD simulations are shown for toxin A i.e., 1DUA (Black) and toxin B i.e.1DT2 (blue).

The flexibility of different segments of the protein is also revealed by looking at the root mean-square fluctuation (RMSF) of each residue from its time-averaged position.

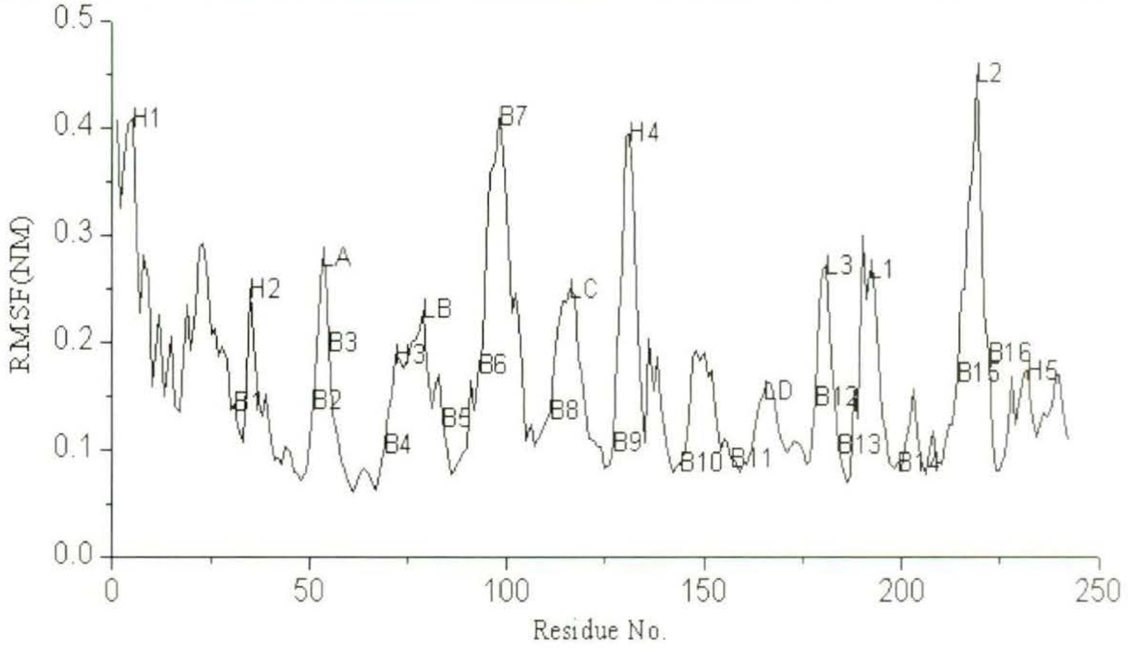


Figure-3a: **Fluctuations:** Root mean squared fluctuations (RMSF) of the Cα atoms during the MD simulations are shown for toxin A (**B**=beta sheet: **H**=helix: L=loop)

It is clear from Fig. 3a and 3b that the loop regions have high fluctuations in case of both toxin A and B. Among the secondary structure beta strand has higher fluctuation than alpha helix. Beta sheet 7 has highest fluctuation and sheets 4, 5, 10, 11 have less fluctuation in case of Toxin A whereas in case of toxin B beta sheet 14 has highest fluctuation and sheets 8, 9, 10 have lower fluctuation. Helix4 shows considerable fluctuations for toxin A and Helix2 and Heilix5 shows considerable fluctuations for toxin B.

In order to further explore the nature of the fluctuations, principal component analysis (PCA) is carried out for both the toxins. (Amadei *et al.,* 1993; Garcia 1992; Das & Mukhopadhyay 2007). It is seen that for the toxin A 74.79% fluctuations are captured by first 10 eigenvectors and the first eigenvector corresponds to 35.49% of the total motion and the second 13.81% and third to a further 8.48 %. On the other hand PCA of the toxin B simulation reveals that the first 10 eigenvectors account for 76.61% of the global motion and that the first eigenvector corresponds to 30.42% of the total motion and the second 19.01% and third to a further 10.73 % The first three eigenvectors with largest eigenvalues were selected as the three principal components PC1, PC2, PC3.

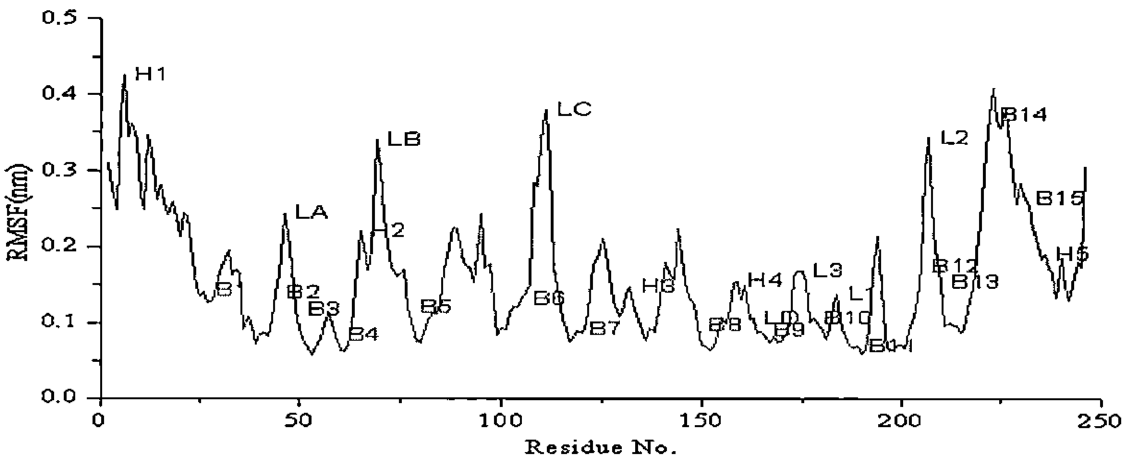


Figure-3b: **Fluctuations:** Root mean squared fluctuations (RMSF) of the Cα atoms during the MD simulations are shown for toxin B (**B**=beta sheet: **H**=helix: **L**=loop)

It is seen that nature of fluctuation of PC1 (Fig. 4a) differs considerably for toxin A and toxin B where as nature fluctuation of PC2 (Fig. 4b) is similar in case of both toxins.

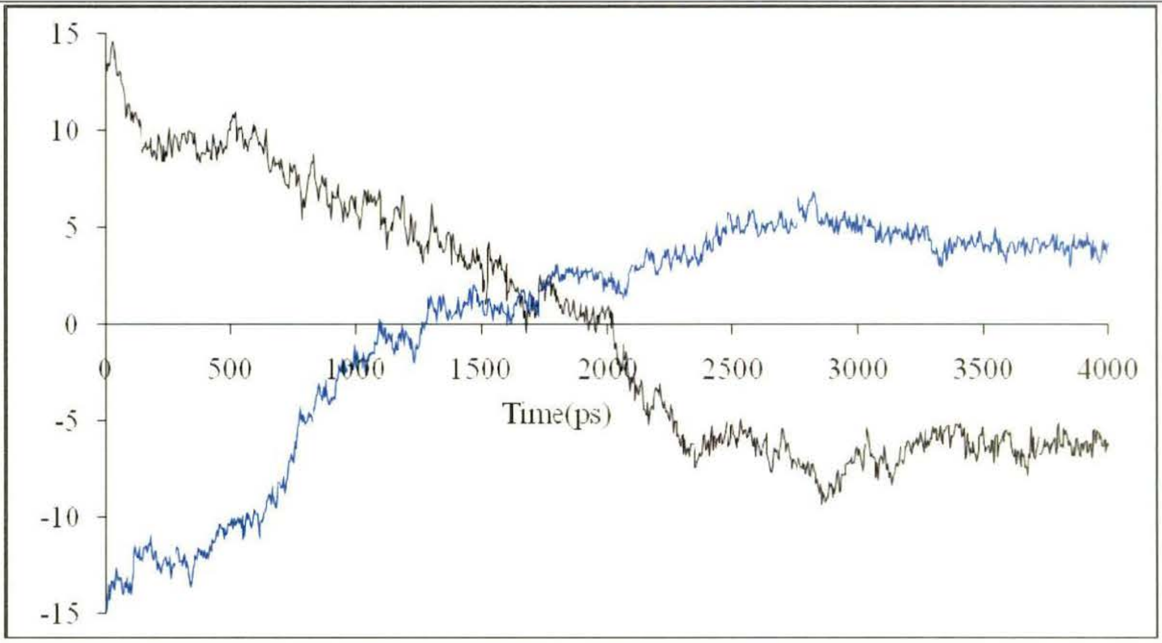


Figure-4a: **Time evolution of PC1**: Variation of principle components 1 with simulation time for toxin A and toxin B[1DUA (Black) and toxin B i.e.1DT2 (blue).

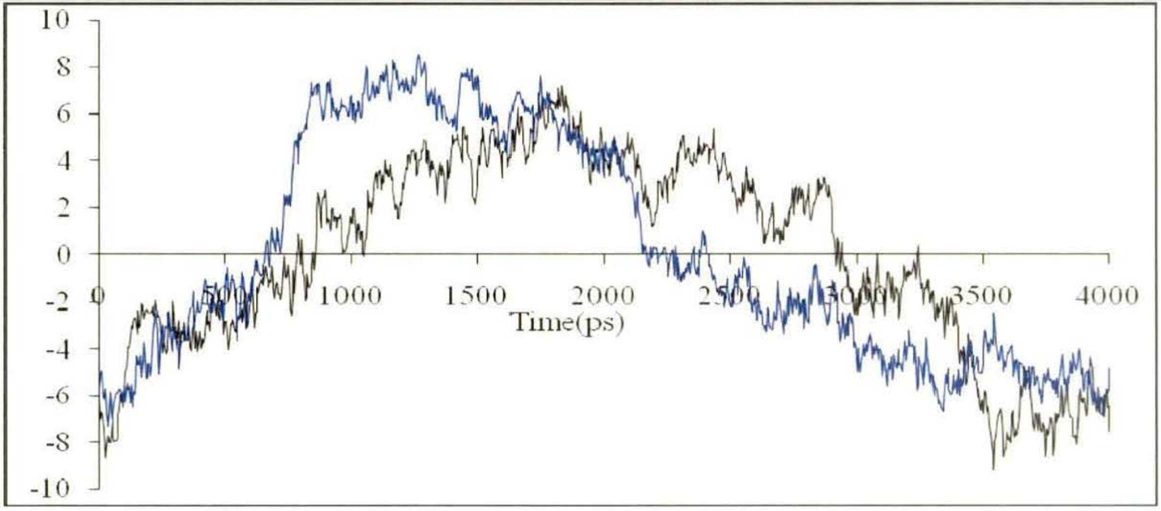


Figure-4b: **Time evolution of PC2**: Variation of principle components 2 with simulation time for toxin B

The RMSF of Cα atoms calculated after projecting trajectories along their respective PC1 and PC2 directions of both the toxins are represented in Fig. 5a and 5b. It is evident for both toxin A and B that fluctuation is highest in the projection on PC1.

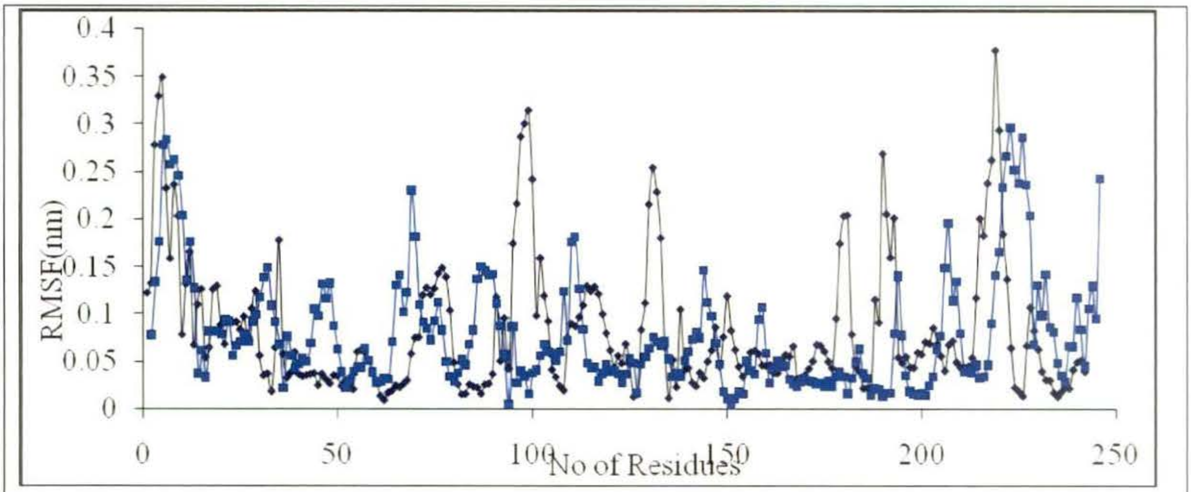


Figure 5a: **Projections on PC1**: The RMSF of Cα atoms calculated after projecting trajectories along their respective PC1 directions [1DUA (Black) and 1DT2 (blue)].

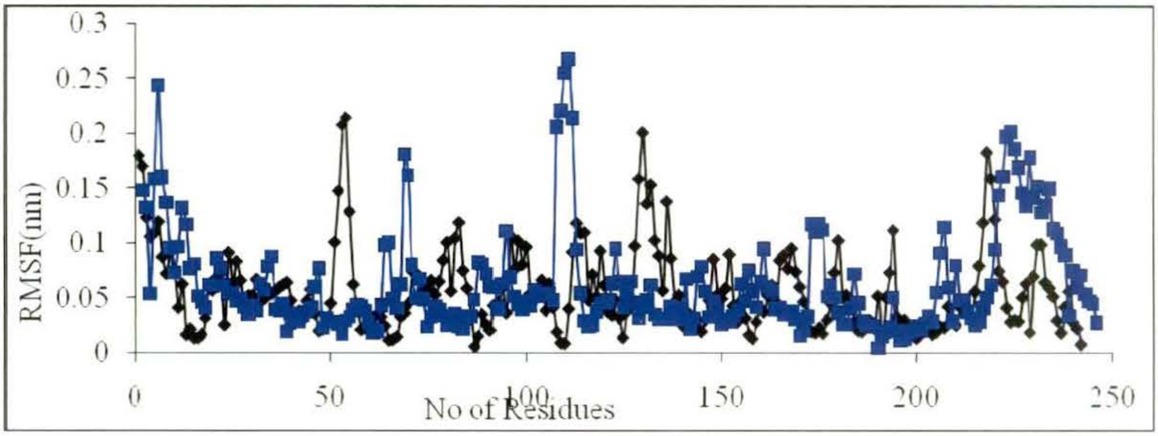


Figure 5b: **Projections on PC2:** The RMSF of Cα atoms calculated after projecting trajectories along their respective PC2 directions [1DUA (Black) and 1DT2 (blue)].

The probability of sampling the phase space determined by first two principal modes during the simulations of toxin A toxin B and presented in Fig. 6.

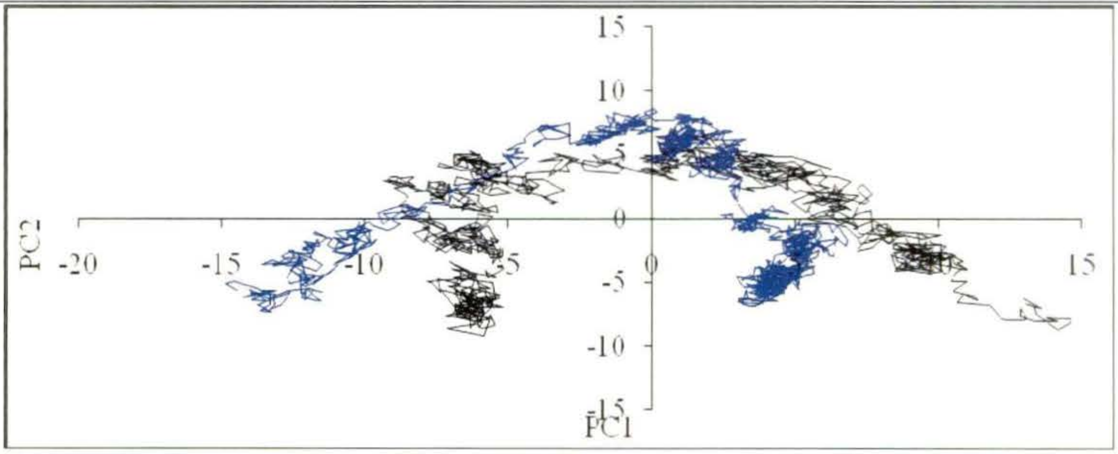


Figure-6: **Conformational Sampling**: The probability of sampling the phase space determined by principal modes 1 and 2 during the simulations of toxin A and toxin B [1DUA (Black) and 1DT2 (blue)].

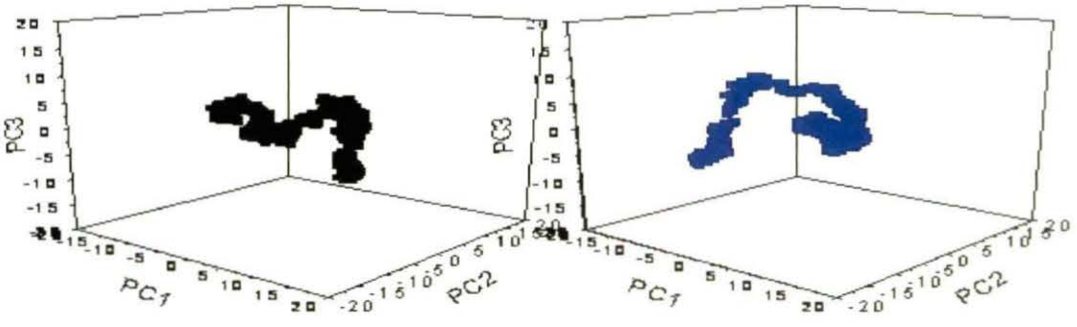


Figure-7: **Space determined PC1, PC2 & PC3 representation**: The probability of accessing regions of the phase space determined PC1, PC2 & PC3 representation for toxin A and B[1DUA (Black) and 1DT2 (blue)].

We further examine the probability of accessing regions of the phase space determined PC1, PC2 & PC3 (Fig. 7) and it is clear that both the toxin show no arrangement in the XY plane and for toxin A almost equally arrange in both the plane YZ and XZ. But molecules of toxin B are more arranged in XZ plane than YZ plane.

In rETA the RMSD values of catalytic residues for Ser195, His72 and Asp120 0.0683, 0.0997, 0.0975 respectively and in rETB, the RMSD values of Ser186, His65, and Asp114 0.038048, 0.1109 and 0.08821. Both Ser195 and Asp120 residue of catalytic triad of rETA are more flexible than in rETB and His65 residue of catalytic triad of rETB is more flexible than in rETA.

In both toxin structures, the primary specificity pocket S1 is formed by residues of domain2 located at the C-terminal end of the molecule in rETA and rETB (in parentheses) are His210 (201), Lys213 (204), Tyr186 (177), and Thr190 (181). In rETA His210 (beta sheet 15), Lys 213(beta sheet 15), Tyr186 (beta sheet 13) and Thr190 (loop1) are important residues for determination the specificity of the substrate and their the RMSD values are 0.0349, 0.073, 0.0273 and 0.0515 respectively and in rETB, the RMSD values of His201 (beta sheet 12), Lys204 (beta sheet 12), Tyr177 (beta sheet 10) and Thr181 (loop1) are 0.031994, 0.074635, 0.028007 and 0.020468. It is clear that only Thr190 (loop1) has high fluctuation in toxin A in comparison to toxin B where the fluctuations of other residues are comparable. Hydrogen bond network is responsible for the overall stability of the protein during simulation although, during whole simulation several hydrogen bonds break and formed.

It is found that the number of hydrogen bonds ranges from 143 to 197 for Toxin B and 141-205 for toxin A. While analyzing the hydrogen bond of binding site it is found only one hydrogen bond remains stable in toxin A throughout the simulation (HB9 formed between carbonyl oxygen of Val119 and the amide nitrogen atoms of Leu121) which belongs to loop C Whereas in case of toxin B three hydrogen bonds remain intact (HB2-, HB4 formed between amide nitrogen of Gly184 and the carbonyl oxygen atoms of Glu182 & the side chain nitrogen atoms of Asn185, HB9 the hydrogen atom of hydroxyl group of Tyr177 and the side chain nitrogen atom of his 201(H9)).HB2 and HB4 form within the Loop1 residues and HB9 forms within beta sheets 10 and 12.

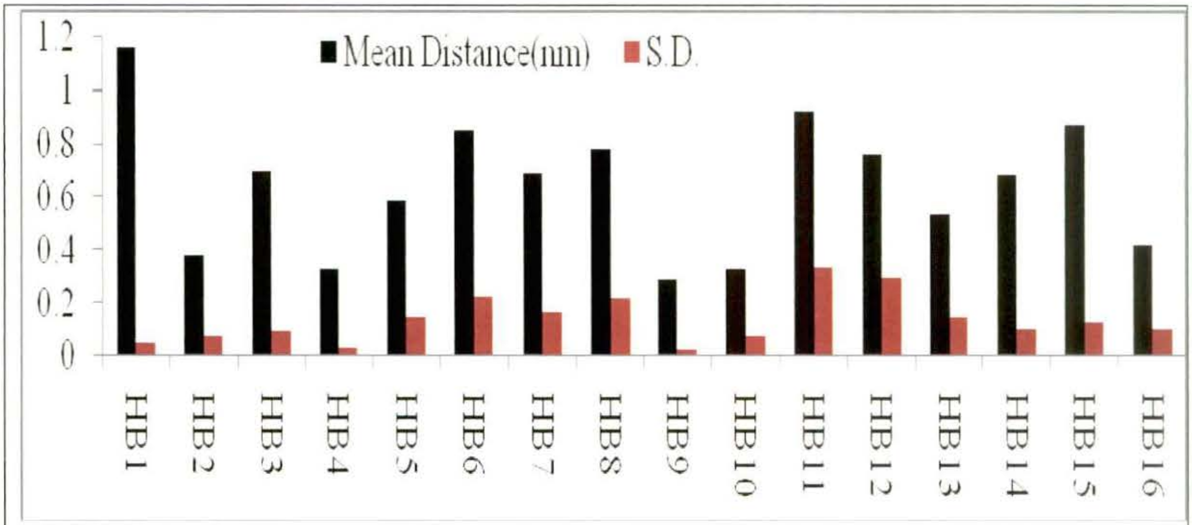


Figure-8a: **Hydrogen bond of toxin A:** Hydrogen bonds within the binding site region of toxin A during the simulation time

HB1: Hydrogen bonds between carbonyl oxygen of Trp14 and the amide nitrogen atom of Tyr18

HB2: Hydrogen bond between side chain oxygen of Tyr18 and the amide nitrogen atoms of Arg71

HB3: Hydrogen bond between the side chain oxygen of Try18 and the amide nitrogen atoms of Thr190

HB4: Hydrogen bond between the carbonyl oxygen of Arg71 and the amide nitrogen atoms of Ala74

HB5: Hydrogen bond between the side chain oxygen of Asp120 and the amide nitrogen atoms of Arg71

HB6: Hydrogen bond between the side chain oxygen of Asp120 and the side chain nitrogen atoms of His72

HB7: Hydrogen bond between the side chain oxygen of Asp120 and the amide nitrogen atoms of His72

HB8: Hydrogen bond between both side chains of Asp120 and Ser211,

HB9: Hydrogen bond between carbonyl oxygen of Val119 and the amide nitrogen atoms of Leu121

HB10: Hydrogen bond between the hydrogen atom of hydroxyl group Tyr186 and the side chain nitrogen atom of His210

HB11: Hydrogen bond between both side chains of Thr190 and Lys213

HB12: Hydrogen bond between the side chain oxygen of Asn194 and the side chain nitrogen atoms of Val191

HB13: Hydrogen bond between the carbonyl oxygen of Val191 and the amide nitrogen atoms of Asn194

HB14: Hydrogen bond between the carbonyl oxygen of Pro192 and the amide nitrogen atoms of Ser195

HB15: Hydrogen bond between the carbonyl oxygen of Pro192 and the side chain of Ser195 (H15),

HB16: Hydrogen bond between the side chain oxygen of Asp164 and the amide nitrogen atoms of Gly193

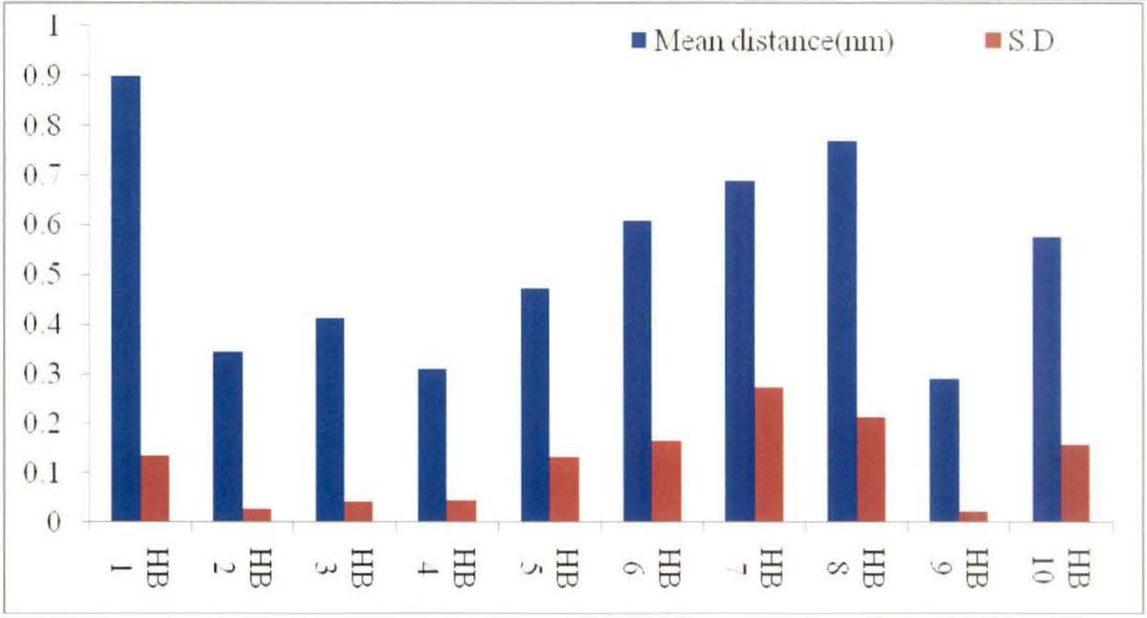


Figure-8b: **Hydrogen bond of toxin B:** Hydrogen bonds within the binding site region of toxin B during the simulation time

HB1: Hydrogen bond between amide nitrogen of Asn158 and the side chain oxygen atoms of Glu182

HB2: Hydrogen bonds formed between amide nitrogen of Gly184 and the carbonyl oxygen atoms of Glu182 (H2)

HB3: Hydrogen bonds between carbonyl oxygen of Glu182 and the amide nitrogen atoms of Asn185 (H3),

HB4: Hydrogen bonds formed between amide nitrogen of Gly184 and the side chain nitrogen atoms of Asn185 (H4),

HB5: Hydrogen bond between the side chain oxygen atom of Glu182 and the hydrogen atom of hydroxyl group Tyr180

HB6: Hydrogen bond between amide nitrogen of Tyr64 and the side chain oxygen atoms of Asp114 (H6)

HB7: Hydrogen bond between the amide nitrogen of His65 and the side chain oxygen atoms of Asp114 (H7),

HB8: side chains of Asp114 and Ser202 (H8),

HB9: Hydrogen bond between the hydrogen atom of hydroxyl group Tyr177 and the side chain nitrogen atom of His201 (H9),

HB10: Hydrogen bond between both side chains of His65 and Ser186,(H10)

There are several hydrogen bonds, which fluctuate during simulation time, some time they cross the limit of hydrogen bond distance (HB2, HB4, HB10, HB16 for toxin A and HB3for toxin B)(Fig. 8a, 8b). Other hydrogen bonds in binding site region for both the toxins were broken during simulation time.

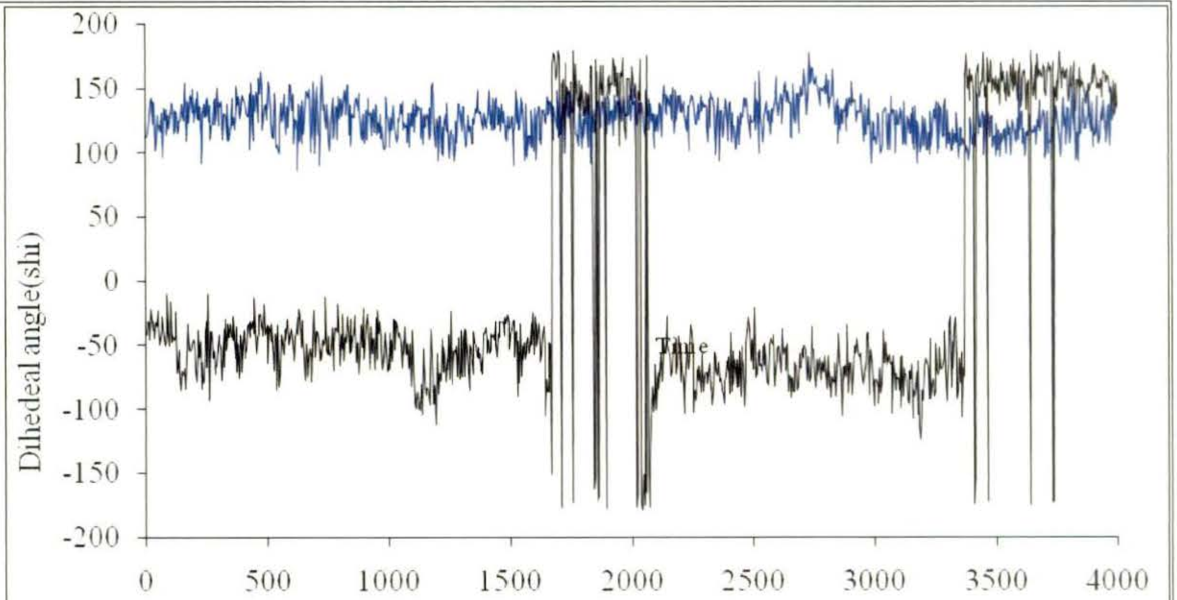


Figure-9: **Dihedral angle of catalytic triad**: Distribution of shi for the Res. Pro192 of toxinA and val183 of toxin B during simulation [1DUA (Black) and 1Dt2 (blue)].

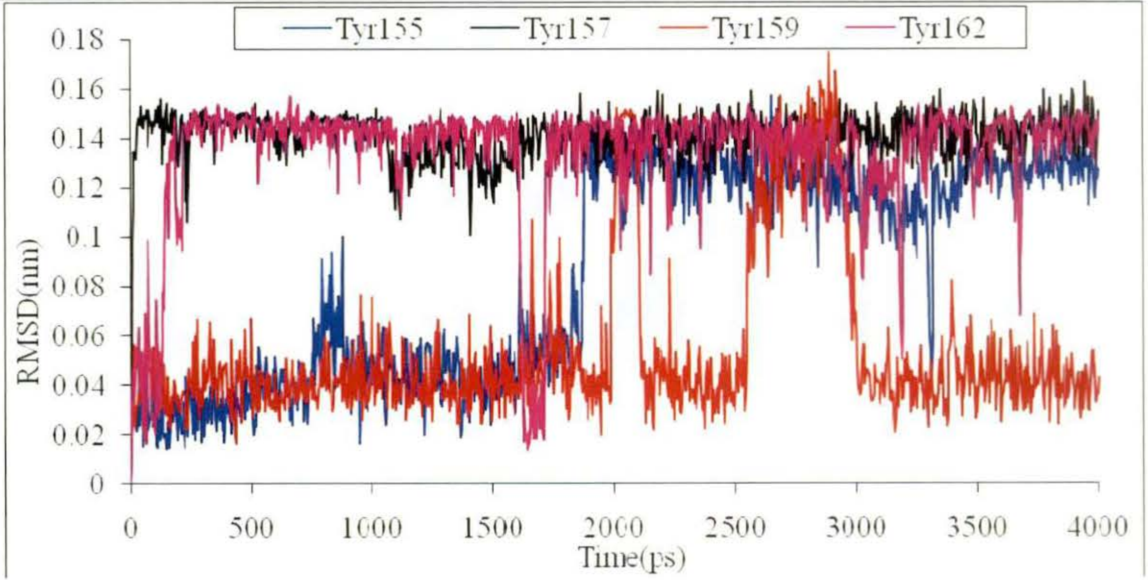


Figure-10: **RMSD of cluster of tyrosine residues:** Time evolution RMSD of cluster of tyrosine residues in rETB (positions 155, 157, 159, and 162)

Time evolution of ψ for the Res. Pro192 of oxyanion hole of toxin A and val183 of toxin B during simulation are shown in Fig. 9. It is seen that Pro192 of toxin A shows a conformational transition where Val183 (corresponding residue from sequence alignment) do not show any conformational transition.

Among the four tyrosine residue it is seen that fluctuation of Tyr162 is lowest and that of Tyr157 and Tyr9 are large indicating their high mobility which enable them to interact with other residues (Fig. 10).

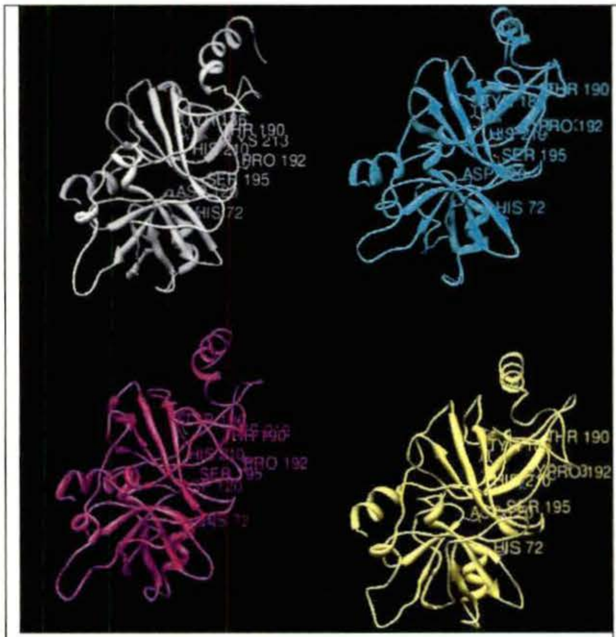


Figure-11a: **Snapshot:** Snapshot of the toxin A (200ps- white 1100ps-pink, 2100ps- green and 3000ps- yellow) are presented in this figure

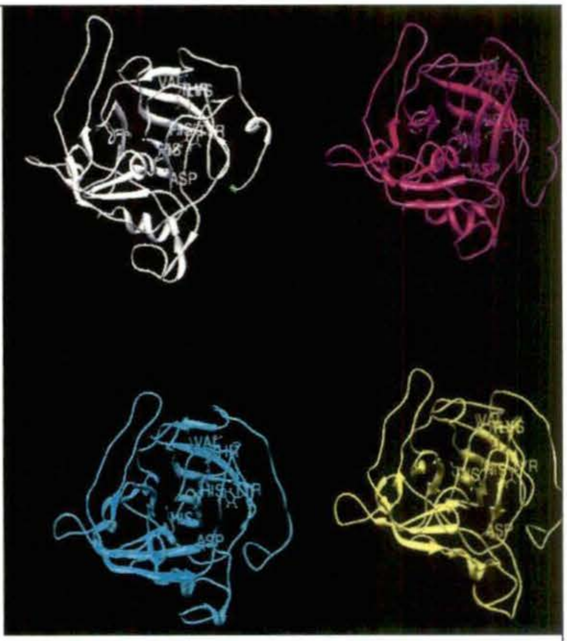


Figure-11b: **Snapshot:** Snapshot of the toxin B (200ps- white 1100ps-pink, 2100ps- green and 3000ps- yellow) are presented in this figure

Snapshot of the toxin are presented in Fig. 11a and 11b where Pro192 (Val183) an important residue of oxyanion hole, catalytic triad residues Ser195 (Ser186), His72 (his65) & Asp120 (Asp114) and the primary specificity pocket S1 residues His210 (His201), Lys213 (Lys204), Tyr186 (Tyr177) & Thr190 (Thr181) are labeled for rETA and rETB respectively (in parentheses).

The plot of distance of Cα-Cα of interacting residues of S1 subsite and helix 4 of toxin B residues during simulation time (Fig. 12) shows the residue come together during simulation and interact with each other, we tried to find whether any hydrogen bond is formed between these residues, we observed no hydrogen bond is formed. So vanderwaals interaction is responsible for their interaction.

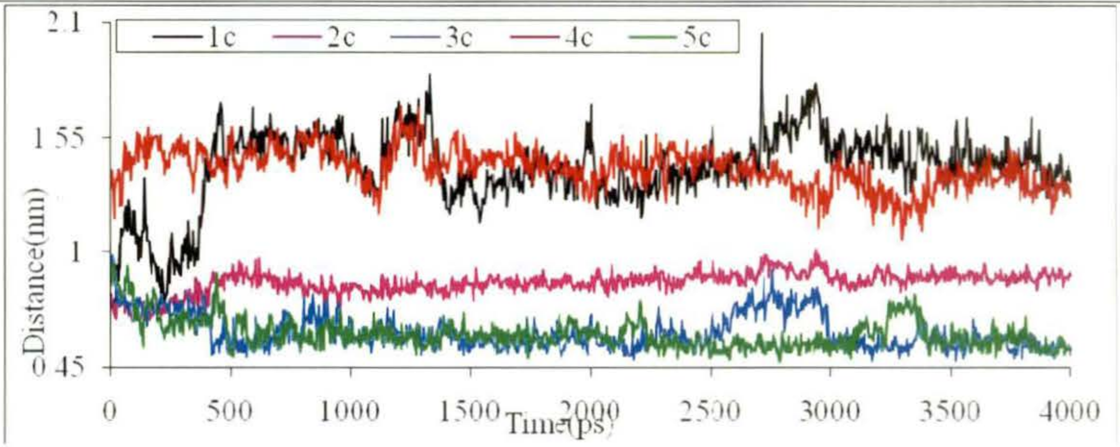


Figure-12: **Cα-Cα distance of interacting residues**: Time evolution Cα-Cα distance of interacting residues of S1 subsite and helix 4 of toxin B

C1: Distance between Cα atoms of residues Pro19 and tyr159

C2: Distance between Cα atoms of residues Glu182 and tyr157

C3: Distance between Cα atoms of residues Val183 and tyr159

C4: Distance between Cα atoms of residues Glu16 and Lys204

C5: Distance between Cα atoms of residues Thr181 and Lys204

7.4: Discussion

Toxin A and B have high degree of sequence and structural similarity. We investigated the motional properties of the toxins in this work. During the dynamics, the structural variations, as measured by the RMSD and the radius of gyration as a function time for both toxins suggest that the simulations are stable. It is also evident that toxin B is somewhat more rigid than toxin A. From RMSF analysis it is evident that nature of fluctuation for different secondary structures and loop regions are similar in both toxin A and toxin B.

From the RMSF of Cα atoms calculated after projecting trajectories along their respective Principal components (PC1 & PC2) indicate that fluctuation is higher in case of toxin A along both the PC i.e. PC1 & PC2. (Fig. 5a, 5b).

It is clear from the probability of sampling the phase space determined by first two principal modes during the simulations the projection of the dynamics trajectory onto the first two PC that the protein A traverse one conformational space around the origin and another at the right side of the origin and also at the left side of the origin. It is evident from Fig. 6 that conformational freedom is more at the left side of the origin. In rETA Hydrogen bond between carbonyl oxygen of Val119 and the amide nitrogen atoms of Leu121 (H9), was very strong suggesting the loop C is less flexible than toxin B. The hydrogen bonds between carbonyl oxygen of Trp14 and the amide nitrogen atom of Tyr18 (H1) were broken during energy minimization which indicates that the helix1 is much flexible.

In rETB Glu182 Gly184 and Asn185 belong to loop1 The two strong hydrogen bonds were observed in the amide nitrogen of Gly184 and the carbonyl oxygen atoms of Glu182 (H2) carbonyl oxygen of Glu182 and the side chain nitrogen atoms of Asn185 (H4), suggest that that loop1, is less flexible in toxin B than toxin A. The hydrogen atom of hydroxyl group of Tyr177 and the side chain nitrogen atom of His 201(H9) forms another strong hydrogen bond also reveal that beta sheets 10 and 12 are also strong because tyr177 belongs beta sheet 10 and the residue His201 belongs from sheet 12. Overall we can say that loop1 and beta sheets of 10 and 12 are conserved. The hydrogen bond between amide nitrogen of Asn158 and the side chain oxygen atoms of Glu182 (H1) were broken during energy minimization increases the flexibility helix-4 because Asn158 belong helix 4, at the same time Glu182 of S1 subsite (loop 1) has less flexible because it forms another two strong hydrogen bonds mention above.

In both toxin structures, the primary specificity pocket S1 is formed by residues of domain 2 and in both structures the binding of substrate in the S1 pocket may be stabilized by His210 (201). (Cavarelli et al. 1997) The positioning of Tyr186 (177) helps in optimal orientation of His210 (201) on one side of the pocket. Moreover, Lys213 (204) a residue conserved in both ETs, interacts with the side chain of the glutamate moiety in the S1 pocket. Low RMSD values of those residues support the above facts. In trypsin-like serine proteases, the cleavage of the scissile bond involves formation of a negative charge that is stabilized in a pocket called the "oxyanion hole," formed by the main-chain atoms (involving the NH groups of Gly193 and Ser195 in ETA and Gly184 and ser 186 in ETB) surrounding the catalytic serine (Papageorgiou *et al.,* 2000). In rETA, Pro192 and Gly193 have a unique conformation. In this case negative charge developed by tetrahedral transition state is stabilized by a conformational change of the loop, which needs a rotation of psi of pro192 from helical to strand nature. Residue Pro192 of Toxin A overcomes the energy barrier and folded into another structure, on the other hand Val183 of toxin B ψ value was very stable, remaining throughout our simulation shown in Fig. 9.

It is seen from the plot of psi against simulation, the dihedral is free to rotation without involvement of much energy. During the simulation several times it jumps from -40^0 regions to 180^0 regions.

Thus inactive toxin A becomes active gaining the preferred conformation. From the time evolution of Val 183 the corresponding residue of toxin B, it is seen that the psi of has very less freedom, its value remains close to 140^0 without any major valine183 deviation. It is clear that in case of toxin B this residue is in preferred conformation to show activity. This observation clearly indicates that pro192 and val183 is key residue towards the activity and why toxin A and toxin B show the difference in initiation of activity.

It is reported that mutation of Tyr157, Tyr159 causes the loss activity of the toxin B (Papageorgiou *et al.,* 2000). High RMSD of these residues is supportive to the fact their interaction S1 subsite residues is energetically favourable. From the plot of distance between Cα-Cα of interacting residues of S1 subsite and helix 4 of toxin B against simulation time it also revealed that these residues are highly mobile and their mobility do not require much energy and any hindrance on their conformational free has immense impact on activity.

7.5: References

Amadei A, Linssen ABM & Berendsen HJC. (1993) Essential dynamics of proteins. *Proteins: Struct. Funct. Genet.* **17**: 412–425.

Arbuthnott JP & Billeliffe B (1976) Qualitative and quantitative methods for detecting staphylococcal epidermolytic toxin. *J. Med. Microbiol.* **9**:191-201.

Bailey CJ & Redpath MB (1992). The esterolytic activity of epidermolytic toxins. *Biochem. J.* **284:**177–180.

Berendsen HJC, Postma JPM, DiNola A & Haak JR. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**:3684–3690.

Cavarelli J, Prevost G, Bourguet W, Moulinier L, Chevrier B, Delagoutte B, Bilwes A, Mourey L, Rifai S, Piemont Y & Moras D (1997) The structure of *Staphylococcus aureus* epidermolytic toxin A, an atypic serine protease, at 1.7 A resolution. *Structure* **5:**813–824.

Darden T, York D & Pedersen L (1993) Particle mesh Ewald: An *N*·log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.* **98**:10089–10092.

Das A & Mukhopadhyay C (2007) Application of principal component analysis in protein unfolding: An all-atom molecular dynamics simulation study. *The Journal of Chemical Physics* **127**, 165103-165108.

Essmann U, Perera L, Berkowitz ML, Darden T, Lee H & Pedersen LG (1995) A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**:8577–8593

Garcia AE (1992) Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* **68**:2696–2699.

Hess B, Bekker H, Berendsen HJC & Fraaije JGE (1997) LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18:**1463-1472.

Jeyakumari D, Gopal R, Eswaran M & MaheshKumar C (2009) The staphylococcal scalded-skin syndrome in a new born. *Journal of Global Infectious Diseases* **1:**45-47.

Johnson-Winegar AD & Spero L (1983) Isoelectric focusing patterns of staphylococcal exfoliative toxin. *Curr. Microbiol.* **8**:311-315.

Kapoor V; Travadi, J & Stephen B (2008) Staphylococcal scalded skin syndrome in an extremely premature neonate: A case report with a brief review of literature. Journal of *Paediatrics & Child Health.* **44:** 374-376

Kapral FA & Miller MM (1971) Product of *Staphylococcus aureus* Responsible for the Scalded-Skin Syndrome. *Infection and Immunity.* **4**: 541-545.

Kondo I, Sakurai S & Sarai Y (1973) Purification of exfoliation produced by *Staphylococcus aureus* of bacteriophage group 2 and its physicochemical properties. *Infect. Immun.* **8**:156-164.

Ladhani S, Joannou CL, Lochrie DP, Evans RW & Poston SM (1999) Clinical, microbial, and biochemical aspects of the exfoliative toxins causing staphylococcal scalded-skin syndrome. *Clin Microbiol Rev.* **12**:224–242.

Lesk AM & Fordham WD (1996) Conservatiuon and variability in the structures of serine proteinases of the chymotrypsin family. *J. Mol. Bio.***258:** 501-537.

Lindahl E, Hess B & van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Modeling* **7**:306-317.

Lowney ED, Baublis JV Kreye GM, Harrell ER & McKenzie AR (1967) The scalded-skin syndrome in small children. *Arch. Dermatology.* **95**:359-369.

Marrack P & Kappler J (1990). The staphylococcal enterotoxins and their relatives. *Science.* **248:**750–711.

Melish ME & Glasgow LA (1970) The staphylococcal scalded-skin syndrome. *N. Engl. J. Med.* **282:**1114–1119.

Monday SR, Vath GM, Ferens WA, Deobald C, Rago JV, Gahr PJ, Monie DD, Iandolo JJ, Stephen K. Chapes SK, Davis WC, Ohlendorf DH, Schlievert PM & Bohach GA (1999). Unique superantigen activity of staphylococcal exfoliative toxins. *J. Immunol.* **162:**4550–4559.

Morlock BA, Spero L & Johnson AD (1980). Mitogenic activity of staphylococcal exfoliative toxin. *Infect. Immun.* **30:**381–384.

Papageorgiou AC, Plano LR, Collins CM & Acharya KR (2000). Structural similarities and differences in *Staphylococcus aureus* exfoliative toxins A and B as revealed by their crystal structures. *Protein Sci.* **9: 610-618.**

Perona JJ & Craik CS (1995) structural basis of substrate specificity in the serine proteases. *Protein Sci.* **4 337-360**

Scopes JW, Eykyn S & Phillips I (1974) Staphylococcal infection in the newborn. *The Lancet* **304**:1392.

Vath GM, Earhart CA, Rago JV, Kim MH, Bohach GA, Schlievert PM & Ohlendorf DH (1997). The structure of the superantigen exfoliative toxin A suggests a novel regulation as a serine protease. *Biochemistry.* **36:**1559–1566.

Vath GM, Earhart CA, Monie DD, Iandolo JJ, Schlievert PM & Ohlendorf DH (1999). The crystal structure of exfoliative toxin B: a superantigen with enzymatic activity. *Biochemistry* **38:**10239–10246.

Wiley BB & Rogolsky M (1977) Molecular and serological differentiation of staphylococcal exfoliative toxin synthesized under chromosomal and plasmid control. Infect. *Immun.* **18**:487-494.

Chapter VIII

Homology modeling and MD simulation of the Cytolethal distending toxin B gene of Helicobacter hepaticus ATCC 51449

8.1: Introduction

H. hepaticus is a naturally occurring pathogen of mice, has been used as a model for the study of hepatic carcinogenesis and gastrointestinal disease. *H. hepaticus* produces a soluble toxin known as cytolethal distending toxin (Cdt). Cdt is a bacterial toxin that induces cell cycle arrest of cultured cells in the G2 phase. It has been found in a number of mucosal pathogens, including *Campylobacter jejuni* (Johnson & Lior 1988) and other *Campylobacter* species (Pickett *et al.,* 1996), certain *Escherichia coli* strains (Bouzari & Varghese 1990; Johnson & Lior 1988b) *Shigella dysenteriae* (Okuda *et al.,* 1995), *Haemophilus ducreyi* (Cope *et al.,* 1997) and *A. actinomycetemcomitans* (Sugai *et al.,* 1998) and *H. hepaticus* (Young *et al.,* 2000). Cdt is composed of three subunits, CdtA, CdtB and CdtC, which form a tripartite complex (Saiki *et al.,* 2001; Saiki *et al.,* 2004; Lara-Tejero & Gala'n 2002). CdtA and CdtC are required for the delivery of CdtB, the active subunit (Lara-Tejero & Gala'n 2001; Deng & Hansen 2003; Lee *et al.,* 2003; Shenker *et al.,* 2004). On delivery into host cells by CdtA and CdtC, the active subunit CdtB is transported to the nucleus and causes DNA damage (Elwell & Dreyfus 2000; Lara-Tejero & Gala'n 2000). CdtB has an amino acid sequence similar to the DNase I family of proteins. The CdtB gene is found to be the most conserved among the three Cdt genes (Pickett et al., 1996; Young *et al.,* 2000; Mayer *et al.,* 1999). Little work has been carried out on the CDT of *H. hepaticus*, owing to a limited toxin production by the bacterium and to the difficulties in producing a recombinant toxin. It makes it difficult to produce the toxin in sufficient quantities to perform cell culture studies (Avenaud *et al.,* 2004).Studies with CDT of *Campylobacter jejuni* and *E. coli* have shown that CdtB has a DNase activity in vitro and suggested that CdtB could be the active subunit of the holotoxin (Lara-Tejero & Galan 2000; Elwell & Dreyfus 2000; Elwell *et al.,* 2001). Indeed, mutations in the CdtB's DNase catalytic site abolish the cellular toxic activity of the holotoxin (Lara-Tejero & Galan 2000; Elwell & Dreyfus 2000). As CdtB DNase activity is very poor in vitro (Lara-Tejero & Galan 2000) some authors suggested that CdtB could be a phosphatase for either Weel kinase or CDC25 phosphatase, or for other cell cycle regulatory proteins (Pickett & Whitehouse 1999; Dlakic 2000; Dlakic 2001).

Cdt induces DNA double strand breaks in target cells. This was shown in yeast cells transfected with the CdtB gene (Hassane *et al.*, 2001) and in mammalian cells naturally intoxicated by extracellular treatment with the toxin (Frisan *et al.*, 2003). The subcellular target of Cdt is known to be DNA. In *H. hepaticus* ATCC51449, four conserved residues (S93, R139, N190 and T224) are present (Dassanayake *et al.*, 2005).

It is observed that the CdtB does indeed possess a DNase activity, which is absent in the CdtA and CdtC subunits, and that this activity can be abolished by an amino acid change in the region identified as potentially responsible for this activity. Three CdtB mutants of *H. hepaticus* were constructed by Avenaud *et al,* one of them in the potential DNase activity region (CdtBH265L mutant) and two others in the potential phosphatase activity region (CdtBN27I and CdtBE60V mutants) (Avenaud *et al.*, 2004). N27 residue was noted to be important as it surrounds the catalytic pocket and possibly the phosphate (Dlakic 2000). It is now widely documented in several bacteria that CdtB is responsible for the cell toxicity of the CDT via its DNase activity (Lara-Tejero & Galan 2002).

H. hepaticus ATCC51449 was isolated from liver tissue in the course of the initial investigation of the outbreak of hepatitis (Ward *et al.*, 1994). In spite of increasing research effort and complete genome of the bacteria *H. hepaticus* ATCC51449 is sequenced (Suerbaum *et al.*, 2003) but the structure of Cdt of *H. hepticus* is not solved yet.

The aim of this work is to construct the 3D model of CdtB proteins from *H. hepaticus* ATCC51449 strains taking the crystal structure of Cytolethal distending toxin B (CdtB) of

A. actinomycetemcomitans (Yamada *et al.,* 2006) as template and to perform molecular dynamics simulation and principal component analysis to understand its motional properties and conformational space traversed by it.

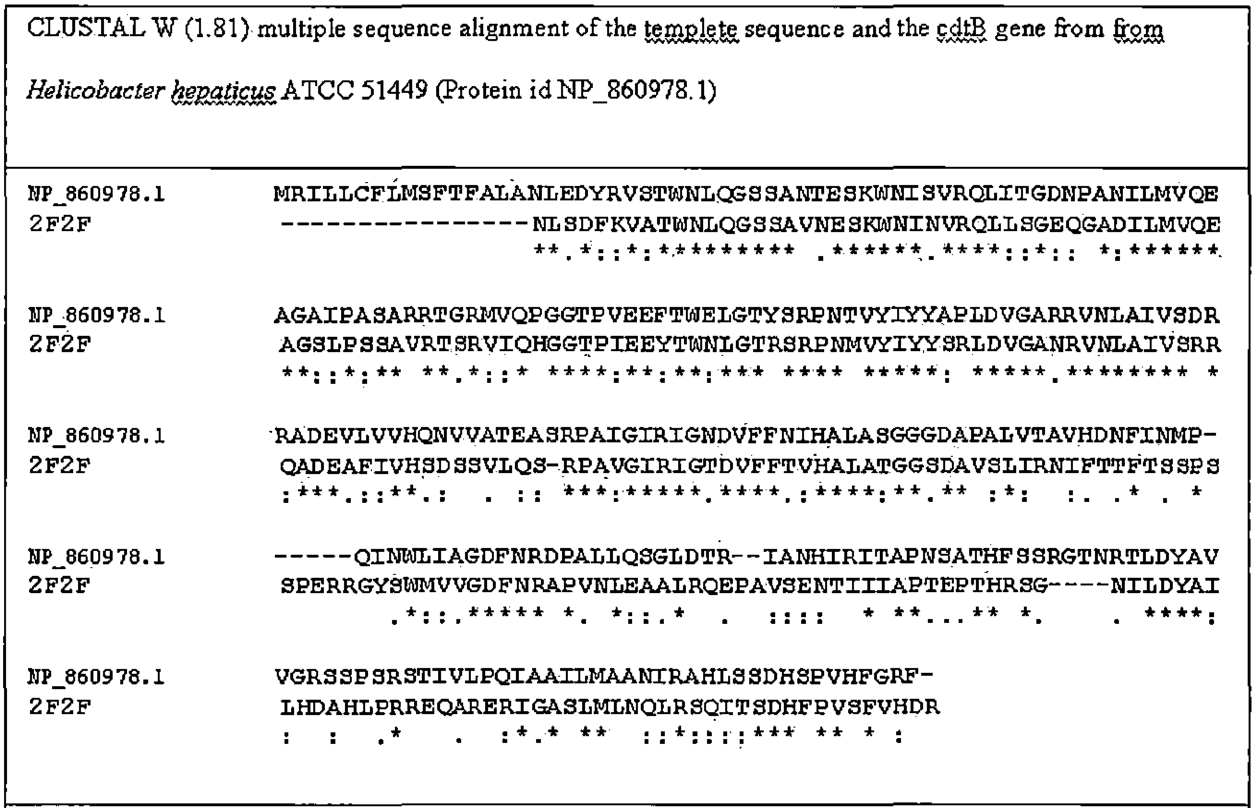
8.2: Materials and methods

The amino acid sequences of the Cytolethal distending toxin B gene (CdtB) of *H. hepaticus* ATCC 51449 bearing protein id NP_860978.1 was obtained from the IMG database (www.img.jgi.doe.gov) (Suerbaum *et al.,* 2003).The protein is 273 amino acid length and it was confirmed that the 3D structure of the protein was not available in Protein Data Bank (http://www.rcsb.org/pdb/home/home.do), consequently the current work of constructing the 3D model of the CdtB gene of *H. hepaticus* ATCC 51449 strains was initiated.

Homology modeling

The preliminary task in the homology modeling technique is to recognize protein structures linked to the target sequence and subsequently select those that will be used as templates (Centeno et al., 2005). PSI-BLAST (Altschul et al., 1997) was carried out against database specification of PDB proteins, which were available at the National Centre for Biotechnology Information (NCBI) Web server (http://www.ncbi.nlm.nih.gov/blast/) to find out remote similarities. We retrieved the known homologous structures of CdtB of *H. hepaticus* ATCC 51449 from the protein data Bank (PDB). We found that protein Cdt from *A. actinomycetemcomitans* [PDB entry: 2F2F (Chain B)] was structural template (percentage sequence identity 47.12) (See Fig. 1).

An optimal alignment between the target sequence and template is required to construct a 3D model of the target protein, after the template sequence has been recognized. Multiple sequence alignments were performed using ClustalW 1.83 using default settings and the aligned sequences were extracted in (.) PIR format (Thompson *et al.,* 1994). The aligned sequences were converted into (.) ALI format (Sali & Blundell 1993) the acquired alignments were crucially assessed in terms of number, length and position of the gaps to make it more reliable.



CLUSTAL W (1.81) multiple sequence alignment of the templete sequence and the cdtB gene from from *Helicobacter hepaticus* ATCC 51449 (Protein id NP_860978.1)

```
NP_860978.1      MRILLCFLMSFTFALANLEDYRVSTWNLQGSSANTESKWNISVRQLITGDNPANILMVQE
2F2F             ----------------NLSDFKVATWNLQGSSAVNESKWNINVRQLLSGEQGADILMVQE
                                 **.*::*:*.********* .******.****::*:: *:******

NP_860978.1      AGAIPASARRTGRMVQPGGTPVEEFTWELGTYSRPNTVYIYYAPLDVGARRVNLAIVSDR
2F2F             AGSLPSSAVRTSRVIQHGGTPIEEYTWNLGTRSRPNMVYIYYSRLDVGANRVNLAIVSRR
                 **::*:** **.*::* ****:**:**:*** **** *****: *****.******** *

NP_860978.1      RADEVLVVHQNVVATEASRPAIGIRIGNDVFFNIHALASGGGDAPALVTAVHDNFINMP-
2F2F             QADEAFIVHSDSSVLQS-RPAVGIRIGTDVFFTVHALATGGSDAVSLIRNIFTTFTSSPS
                 :***.::**.:   .  :: ***:*****.****.:****:**.** :*:  :. .* . *

NP_860978.1      -----QINWLIAGDFNRDPALLQSGLDTR--IANHIRITAPNSATHFSSRGTNRTLDYAV
2F2F             SPERRGYSWMVVGDFNRAPVNLEAALRQEPAVSENTIIIAPTEPTHRSG----NILDYAI
                      .*::.***** *. *::.*  .  ::::  * **...** *.     . ****:

NP_860978.1      VGRSSPSRSTIVLPQIAAILMAANIRAHLSSDHSPVHFGRF-
2F2F             LHDAHLPRREQARERIGASLMLNQLRSQITSDHFPVSFVHDR
                 :  :  .*   .  :*.* **  ::*::::*** ** * :
```

Figure-1: **Alignment:** Alignment of target protein (2F2F) and cytolethal distending toxin protein B (CdtB) from *H. hepaticus* ATCC51449.

The initial 3D model of the CdtB gene of *H. hepaticus* ATCC 51449 was constructed by MODELLER 9v4 program (Sali & Blundell 1993) using the alignment between the CdtB gene of *H. hepaticus* ATCC 51449 and the template protein(2F2F chain B).

Model evaluation:

VERIFY3D was used to validate the refined structures. For evaluation of their internal quality and reliability, the refined model was subjected to the following tests: ProSA (Wiederstein & Sippl 2007) analysis was performed to assess the accuracy and reliability of the modelled structure and check the 3D model for potential errors. Here, the 3D structure of the protein model is compared to its own amino-acid sequence taking into consideration a 3D profile calculated from the atomic coordinates of the structure of correct protein (Eisenberg *et al.,* 1997). Presence of pockets in the structure was predicted using CASTp server (Dundas et al 2006). The refined model was submitted to ProFunc (http://www.ebi.ac.uk/thronton-srv/databases/ProFunc) (Laskowski *et al.,* 2005) to recognize the functional region in the protein. The program PROCHECK (Laskowski *et al.,* 1993) has performed assessment of the predicted model of the CdtB of *H. hepaticus* ATCC 51449 to evaluate their backbone conformation using a Ramachandran plot (Ramachandran *et al.,* 1963). Molecular surface and electrostatic potential of modelled CdtB subunit from *H. hepaticus* ATCC 51449 were predicted using AutoDock Tools version 1.5.2 revision 2 (Goodshell *et al.* 1996).

Molecular dynamics simulation

An understanding of the structural dynamics of the protein is essential to gain greater insights into their important biological functions (Yang *et al.,* 2006). The studies on the structural dynamics were performed using the GROMACS. Taking the rough 3D model of the Cytolethal distending toxin B gene was constructed by MODELLER 9v4 program was used as starting structures for dynamics of the protein model. A single monomer was solvated with SPC water molecules in a cubic box having edge length of 40A˙. The simulation was performed using GROMACS (Lindahl *et al.,* 2001). The LINCS algorithm was used to constrain all bond lengths (Hess *et al.*, 1997). A cutoff of 0.9 nm for Lennard-Jones interactions was used and the particle mesh Ewald method (Darden et al., 1993; Essmann *et al.,* 1995) was employed to calculate longer-range electrostatic contributions on a grid with 0.12-nm spacing and a cutoff of 1.0 nm. The simulation was conducted at constant temperature (300 K), coupling each component separately to a temperature bath using the Berendsen coupling method (Berendsen et al., 1984). The time step was 2 fs, with coordinates stored after every 4 ps. MD simulation was performed for six nano seconds. Before running simulation, an energy minimization was performed in steepest descent method (converged at 107 steps) and this was followed by 1.0ns of simulation imposing positional restraints on the non-H atoms. The positional restraints were then released and 6 ns production run were obtained and analyzed. Analysis programs from GROMACS were used.

Principal component analysis

Principal component analysis (PCA) (Amadei et al., 1993; Garcia 1992; Das & Mukhopadhyay 2007) was performed with the MD trajectory.

8.3: Results

Homology modeling and Model evaluation:

The target sequence [cytolethal distending toxin protein B(CdtB) of *H. hepaticus* ATCC 51449] compared with more identity and releated family using BLAST search and the best template was found 2F2F2 [PDB ID] chain B, which is a crystal structure cytolethal distending toxin protein B(CdtB) of *A. Actinomycetemcomitans.* The B chains of this protein (2F2F2) revealed percent sequence identity 48.71 with target sequence (Fig. 1).

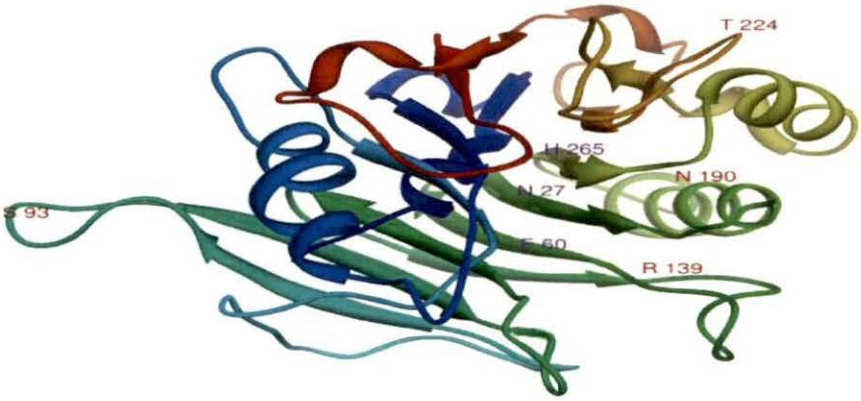


Figure-2: **Ribbon diagram:** The ribbon diagram of the modelled CbtB protein from *H. hepaticus* ATCC51449 is depicted and the four DNA binding residues, namely S93, R139, N190, and T224 are labelled with red colour and the DNase activity region i.e. N27, E60, and H265 are labeled with pink colour.

The modelled structure of the CdtB protein from *H. hepaticus* ATCC51449 is presented in Fig. 2. Four DNA binding residues, namely S93, R139, N190, and T224 are labelled with red colour and the DNase activity region i.e. N27, E60, and H265 are labeled with pink colour in the ribbon diagram of the modelled CbtB protein from *H. hepaticus* ATCC51449 and is also presented in Fig. 2.

Some of the important features of the modelled protein are summarized in Table 1.

Table 1: Summary of the characteristics of the cytolethal distending toxin protein B from *Helicobacter hepaticus* ATCC

General characteristics	CdtB	Characteristics of secondary structure	CdtB
Molecular weight	29,812.8	Alpha helix	20.88%
Net Partial Charge	3.00005	Extended strand	30.77%
Number of atoms	2687	Random coil	48.35%

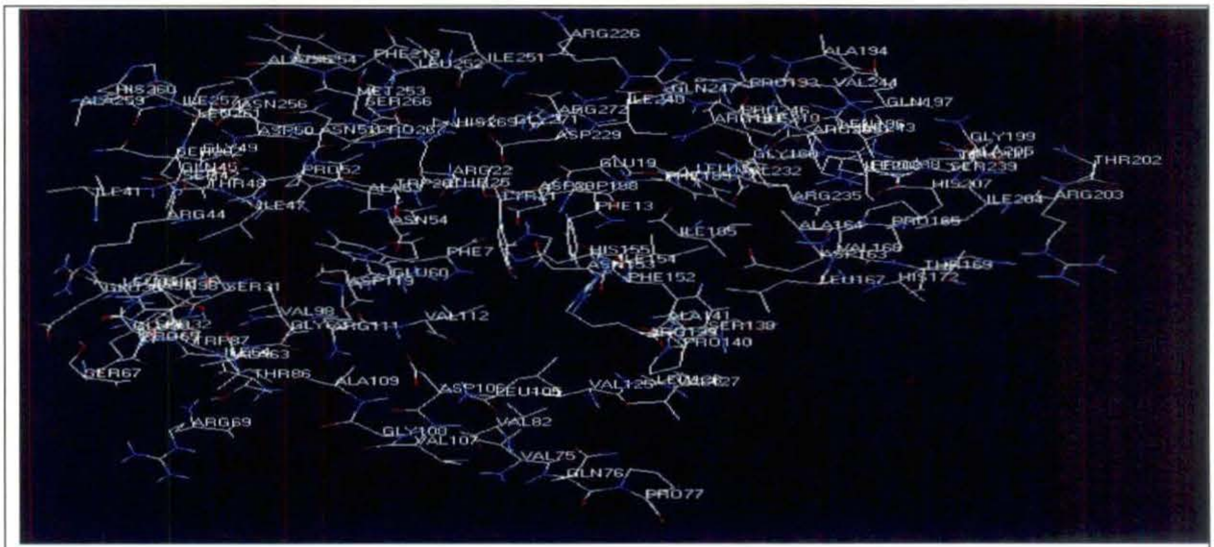


Figure-3: **Pockets for ligands interaction:** Total of 33 pockets for ligands interaction region in the CdtB protein from *H. hepaticus* ATCC51449 are predicted using CASTp servers and visualization on SPDBV software.

CASTp program (Dundas *et al.*, 2006) demonstrated the presence of a total of 33 pockets for ligands interaction in the CdtB protein with varying area and volume (see Table-2).

Table-2: Total of 33 pockets for ligands interaction region in the Cdtb protein from *Helicobacter hepaticus* ATCC51449 is predicted using CASTp server.

Pocket No.	Residue No.	Area	Volume
1	ARG139,PRO140,ILE154,HIS155	24.9	11.6
2	ILE154,ILE167,ILE185	25.5	11.9
3	ARG69,GLU88	0.8	1.7
4	ILE154,ALA164, ILE185,PHE189	25.8	12.2
5	GLY160, ARG191,LEU195,ARG226	24.4	13.7
6	PHE219,HIS260	0.6	1.5
7	ILE41,SER42,ALA259,HIS260	28.4	13.9
8	ALA141,PHE152,ASN153,ILE154,ILE185	29.3	14.6
9	HIS207,ARG235,PRO238,SER239,ILE243	37.1	20.2
10	ALA141,ILE154,LEU167,ILE185	27.4	13.2
11.	THR35,GLU36,ALA63,ILE64,PRO65	26.6	13.2
12	THR25,TRP26,ASP229,SER266	42	23.1
13	ALA254,ILE257,LEU262,SER266,PRO267	35.9	19.6
14.	SER42,GLN45,ILE257,ALA259,HIS260,LEU261	56.1	33.9
15	PHE13,GLU19,TRY21,ILE248,GLY271,ARG272	38	21.4
16.	ARG44,ILE47,THR48,LEU89	21.7	15.8
17	SER138,ARG139,HIS155,LEU157,ASP163	41	20.2
18	GLN197,ILE208,ARG209,ILE210,ILE243,VAL244	39.5	22.3
19	THR169,HIS172,ARG203,ILE204	23.6	14
20	ALA164,VAL168,ILE185,PHE189,LEU196,VAL232	55.6	30.8
21	PRO165,VAL168,GLY199,LEU200,THR202,ARG203,ILE204	64.2	52.3
22	GLU19,ASP20,TYR21,ARG22,GLY271,ARG272	40.3	30.1
23	ASN51,ILE251,LEU252,MET253,ALA255,ASN256,HIS269	50.4	47.1
24	SER67,ARG69,THR86,TRP87,GLU88	64.1	51.3
25	GLU60,ARG111,VAL112,ARG139,HIS155,ALA158	57.7	69.8
26	VAL75,GLN76,PRO77,VAL82,VAL125, LEU126,VAL127	53.3	42
27	GLN197,LEU200,ALA205,ILE208	78.2	64.1
28	PHE7,ILE47,PRO52,ALA53,ASN54,ASN96,VAL98,ASP119	79.7	7.7
29	GLN45,THR48,GLY49,ASP50,ASN51,ASN256	90.6	68.9
30	LEU15,GLU19,ARG235,PRO246,GLN247	67	74.7
31	PRO193,ALA194,GLN197,ILE210	81.5	134
32	SER31,SER32,THR35,GLY62,ALA63,ALA109	106.8	101.3
33	LEU105,ASP106,VAL107,GLY108	130.2	255.6

The constructed model was corroborated by VERIFY 3D (Eisenberg *et al.*, 1997) to estimate the correctness of the model. The z-scores obtained from ProSA analysis for the modelled structures of CdtB protein found to be -5.15.

From the modelled structure using SPDBV program (Guex & Peitsch 1997) residues of 33 pockets for ligands interaction in the CdtB protein are shown in Fig. 3.

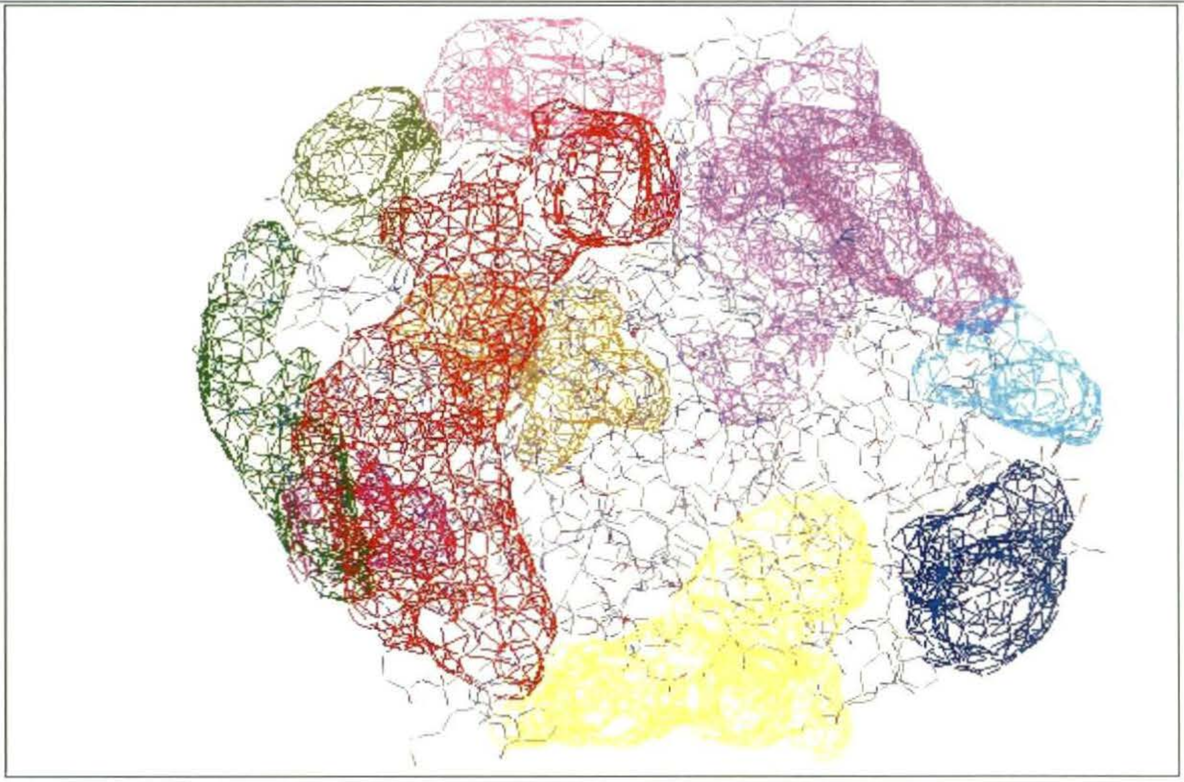
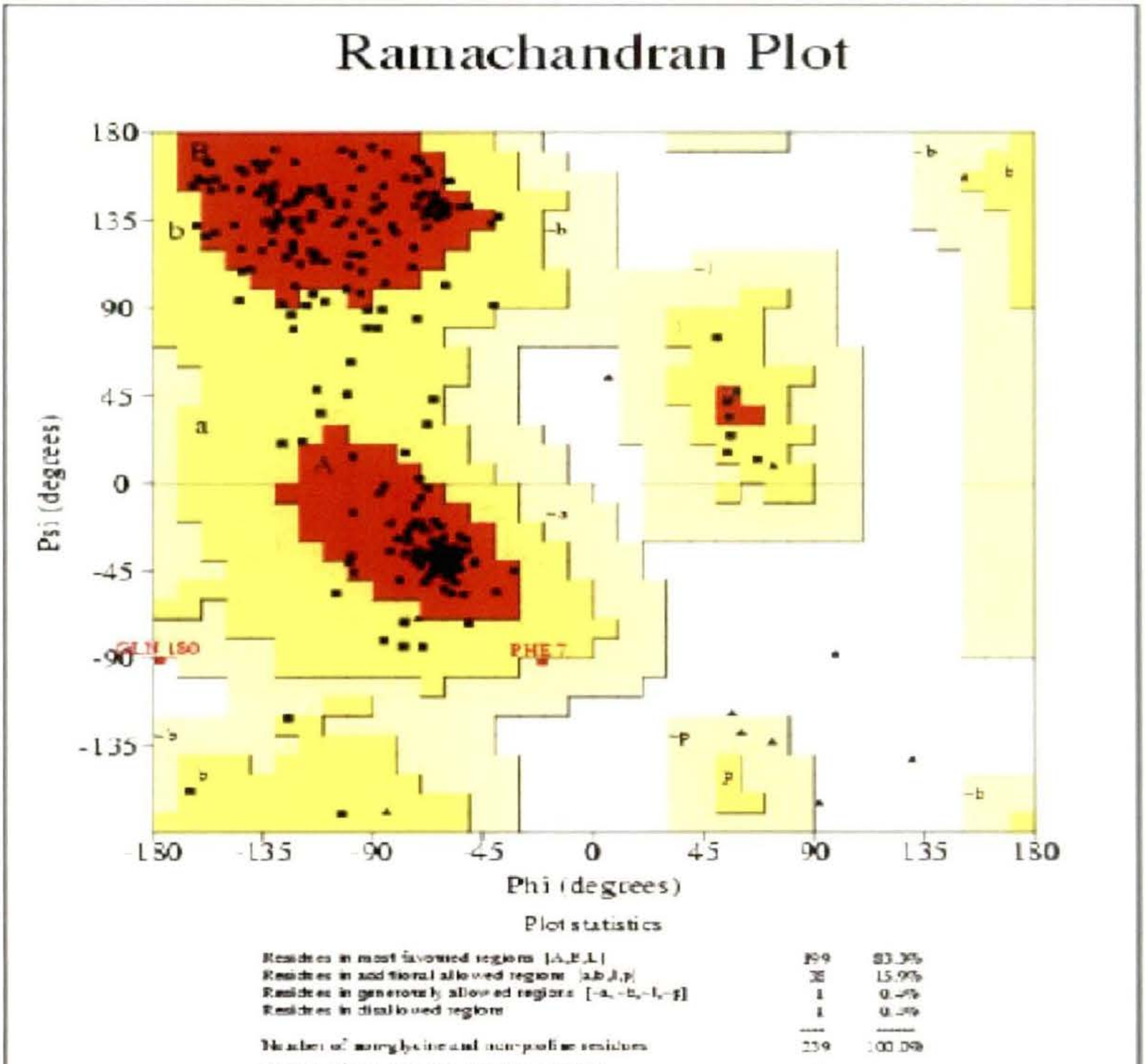


Figure-4: **Clefts and cavities**: Total 10 clefts and cavities in the surface of the protein are predicted using profunc server.

The refined model was submitted to ProFunc (http://www.ebi.ac.uk/thronton-srv/databases/ProFunc) (Laskowski *et al.,* 2005) to analyze the binding sites, showed the presence of 10 clefts and cavities in the surface of the protein shown in Fig. 4.

Nest analysis of the CdtB protein revealed the presence of six nests in this chain. It is found that when modelled CtdB protein from *H.Hepaticus* is compared with different enzyme templets it matches with two Dnase (PDB entry code 1dnk & 2dnj).

The Ramachandran plots (Ramachandran et al., 1963; Rajesh et al., 2007) illustrating the backbone conformation for the modelled protein is presented in Fig. 5.



Plot statistics

Residues in most favoured regions [A,B,L]	199	83.3%
Residues in additional allowed regions [a,b,l,p]	38	15.9%
Residues in generously allowed regions [~a,~b,~l,~p]	1	0.4%
Residues in disallowed regions	1	0.4%
	----	------
Number of non-glycine and non-proline residues	239	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	18	
Number of proline residues	14	

Total number of residues	273	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions

Figure-5: **Ramachandran plot:** Ramachandran plot of the CdtB proteins from *H. hepaticus* ATCC51449 shows the residue backbone conformations for the modelled CdtB protein.

PROCHECK analysis on the stereochemical quality of the 3D model structure of the modelled CdtB protein revealed that 83.3% of residues are in the most favored region in the Ramachandran plot. Moreover, the percentages of residues in the additional allowed regions and generously allowed regions are 15.9% and 0.4%, respectively. However, 0.4% of residues remain on the disallowed region,

Molecular surface and electrostatic potential of modelled CdtB was generated using AutoDock Tools version 1.5.2 revision 2 (Goodshell *et al.*, 1996) and is presented in Fig. 6.

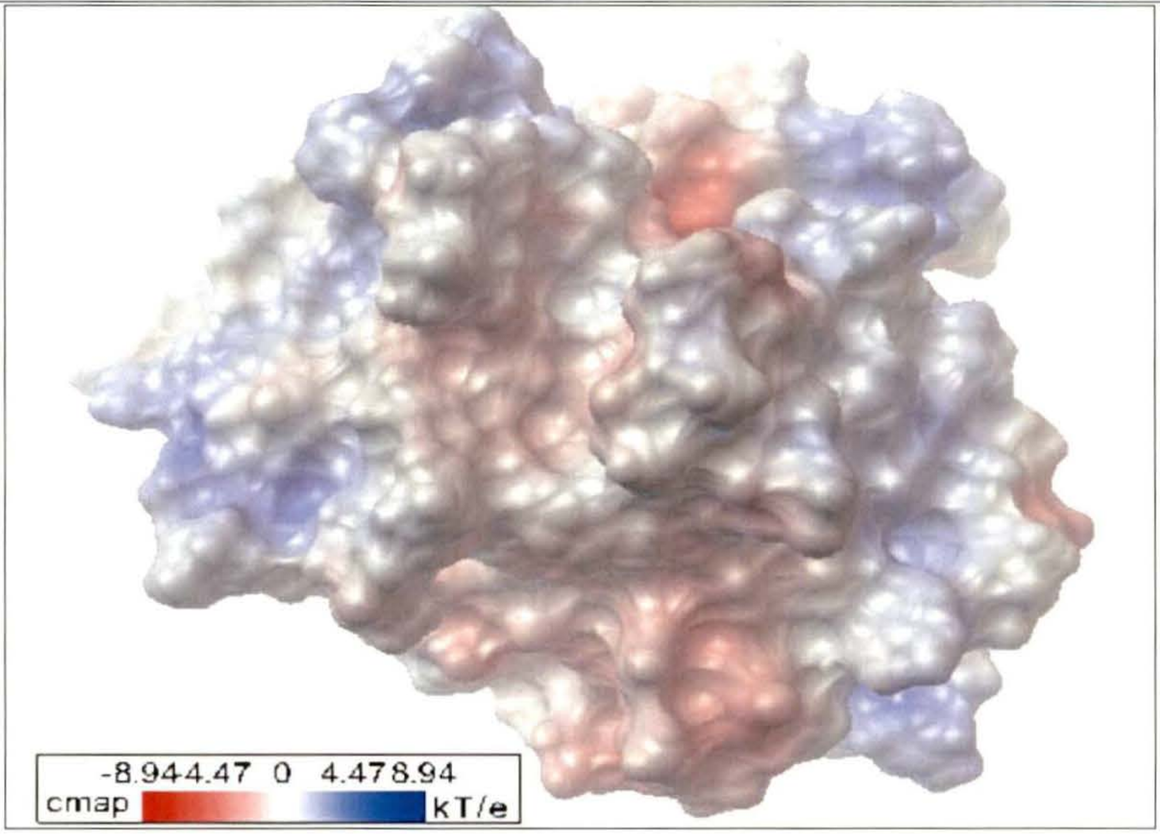


Figure-6: **Electrostatic surfaces:** Electrostatic surfaces of modeled protein CdtB from *H. hepaticus* ATCC51449 was calculated, which positive potential charges in blue, negative values in red and intermediate values in white by using AutoDock tools

Molecular dynamics simulation:

Molecular dynamics simulation of the modelled CdtB is performed and the resulting trajectory is analyzed to study the motional properties of the modelled protein CdtB. The time evolution of root mean square deviation (RMSD) is computed taking the modelled structure (CdtB) of the whole Protein as initial structure and presented in Fig. 7.

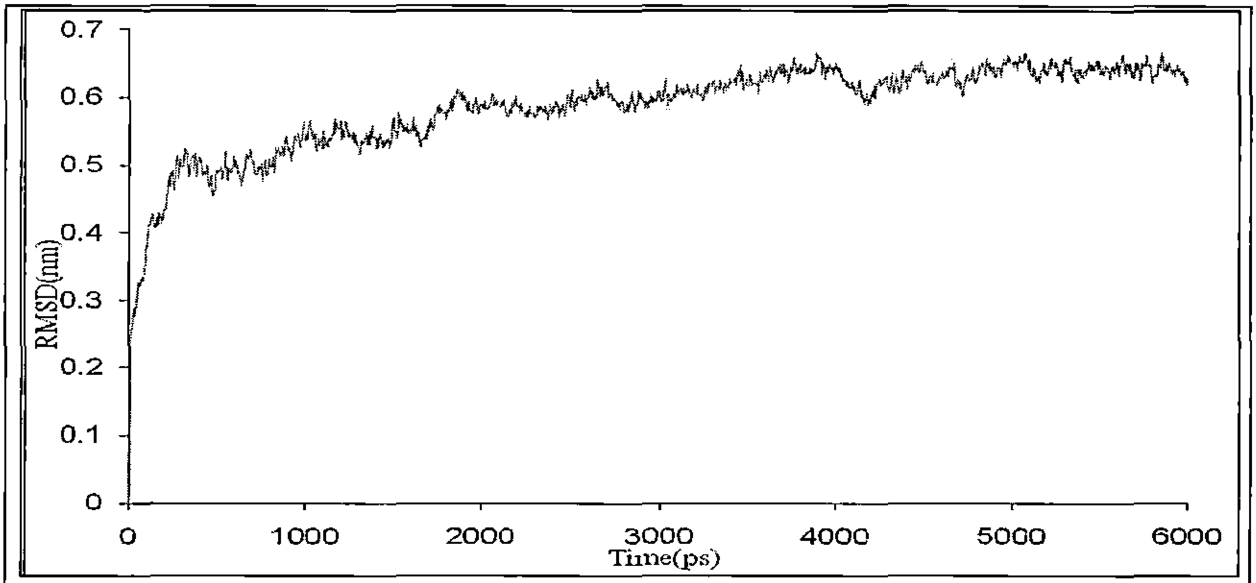


Figure-7: **RMSD:** Root mean square deviations (RMSD) of the CdtB proteins from *H. hepaticus* ATCC51449 protein as a function of time with respect to modeled protein was starting structure during the MD simulations

The time evolution of radius of gyration (Rg) is presented in Fig. 8.

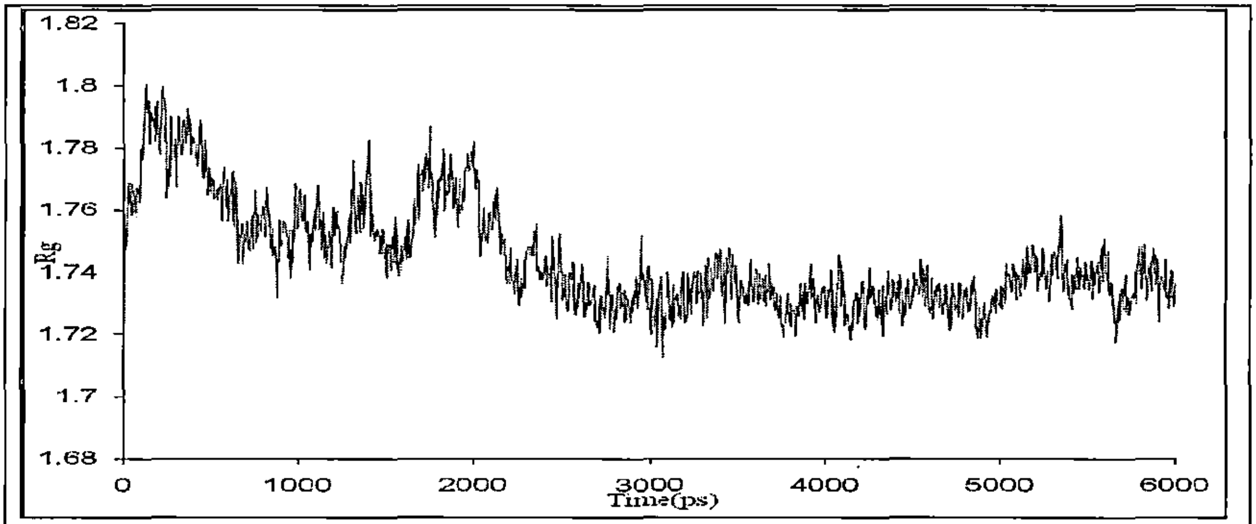


Figure-8: **Radius of gyration**: Radius of gyration (Rg) as a function of time with respect to starting structure during the MD simulations are shown for modelled CdtB protein from *H. hepaticus* ATCC51449

RMSF indicates the flexibility of the protein. RMSF of Cα is presented as a function of residue numbers in Fig. 9.

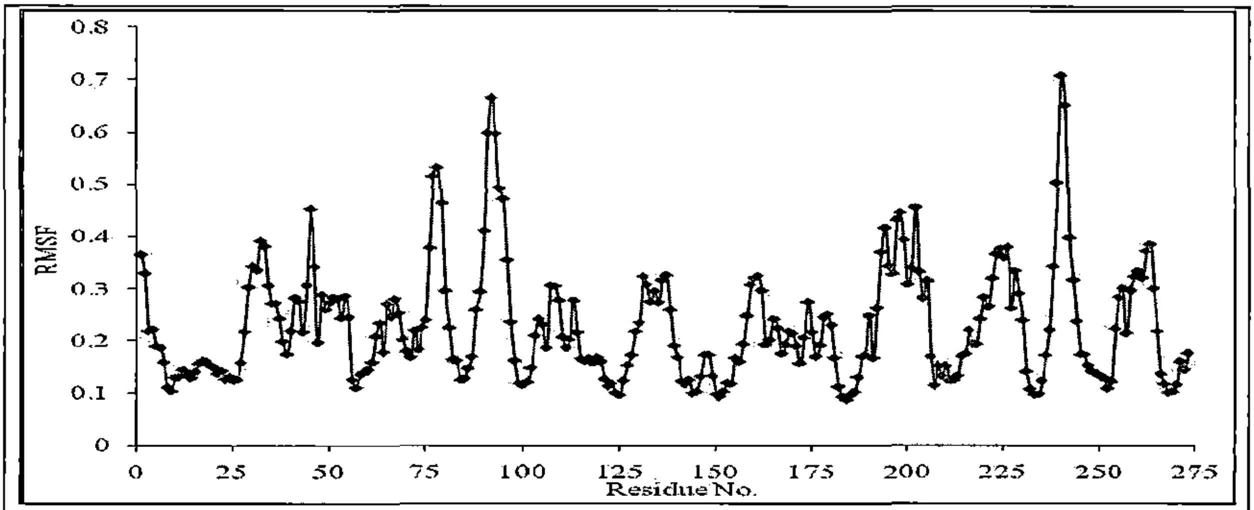


Figure-9: **Fluctuations:** Root means squared fluctuations (RMSF) of the Cα atoms during the MD simulations are shown for modelled CdtB protein of *H. hepaticus* ATCC51449

It is found that the number of hydrogen bonds ranged from 133 to 196 (Fig. 10) during the simulation.

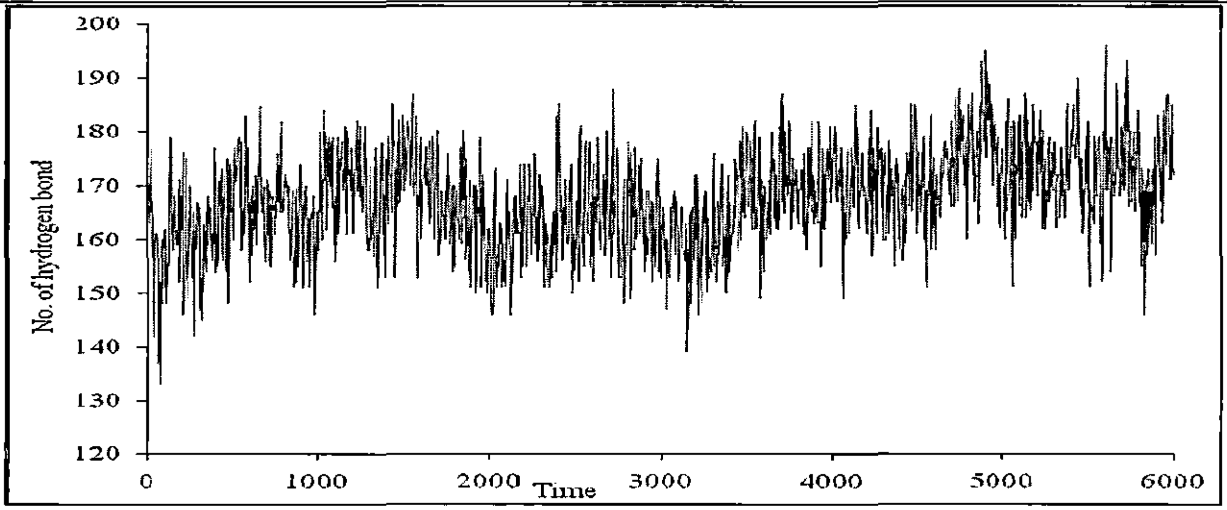


Figure-10: **Hydrogen Bond:** Number of Hydrogen Bonds during the whole Simulation time for modelled CdtB proteins of *Helicobacter hepaticus* ATCC51449

Principal component analysis (PCA): In order to further explore the nature of the fluctuations, principal component analysis (PCA) is carried out for the modelled protein CdtB (Amadei et al., 1993; Garcia 1992; Das & Mukhopadhyay 2007). It is seen that for the toxin 76.12 % fluctuations are captured by first 10 eigenvectors and the first eigenvector corresponds to 40.05 % of the total motion and the second 14.73% and third to a further 5.68 %.

The first four eigenvectors with largest eigenvalues were selected as the four principal components PC1, PC2, PC3 and PC4. Time evolution of principal component 1 (PC1), principal component 2 (PC2), principal component 3 (PC3) and principal component 4 (PC4) in water is represented in Fig. 11.

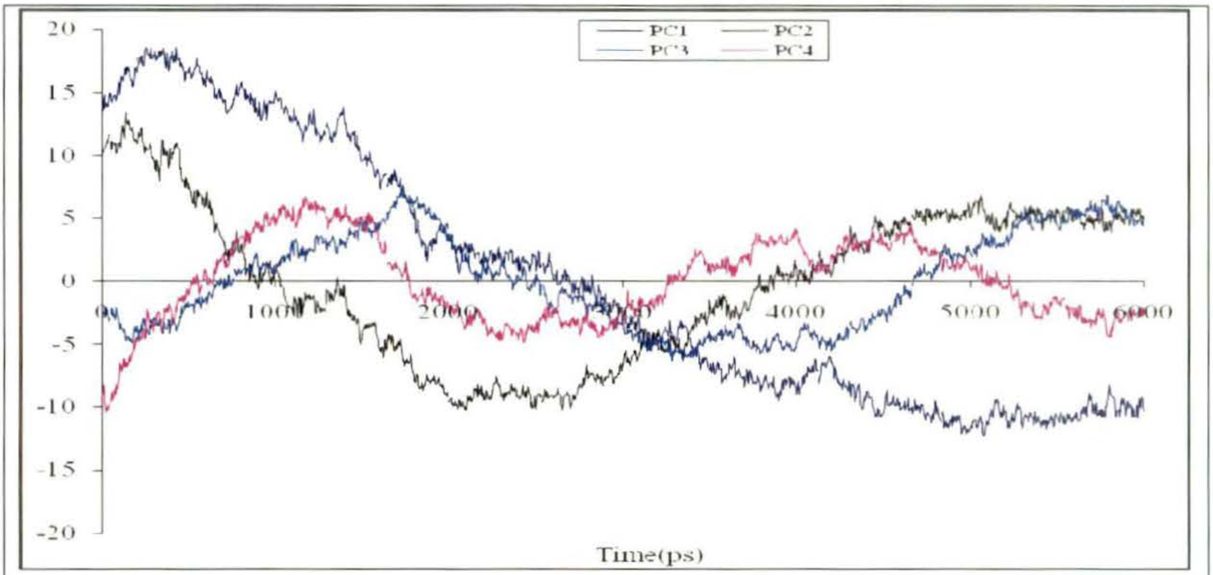


Figure-11: **Time evolution of PC**: Variation of four principle components with simulation time for modelled CdtB protein of *H. hepaticus* ATCC51449

The RMSF of Cα atoms calculated after projecting trajectories along their respective four principal components are presented in Fig. 12. It is evident from the analysis for CdtB that fluctuation is highest in the projection on PC1.

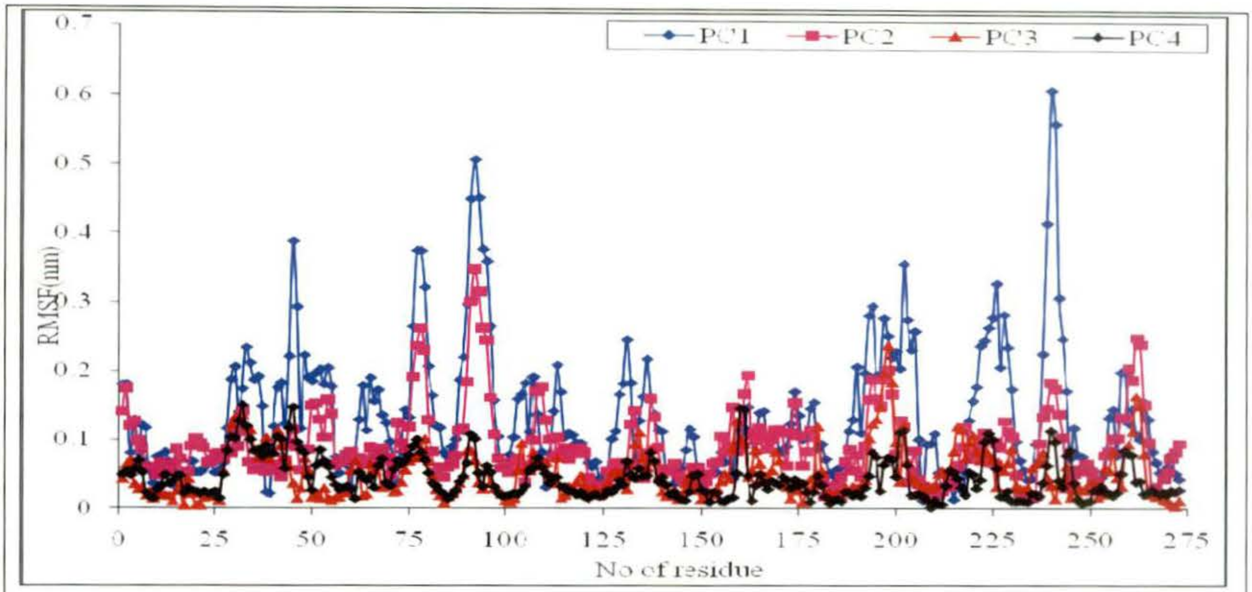


Figure-12: **Projections on PC**: The RMSF of Cα atoms calculated after projecting trajectories along their respective PC1, PC2, PC3 and PC4 directions for the modelled CdtB proteins of *H. hepaticus* ATCC51449

The probability of sampling the phase space determined by the first two principal modes during the simulations of the toxin is presented in Fig. 13.

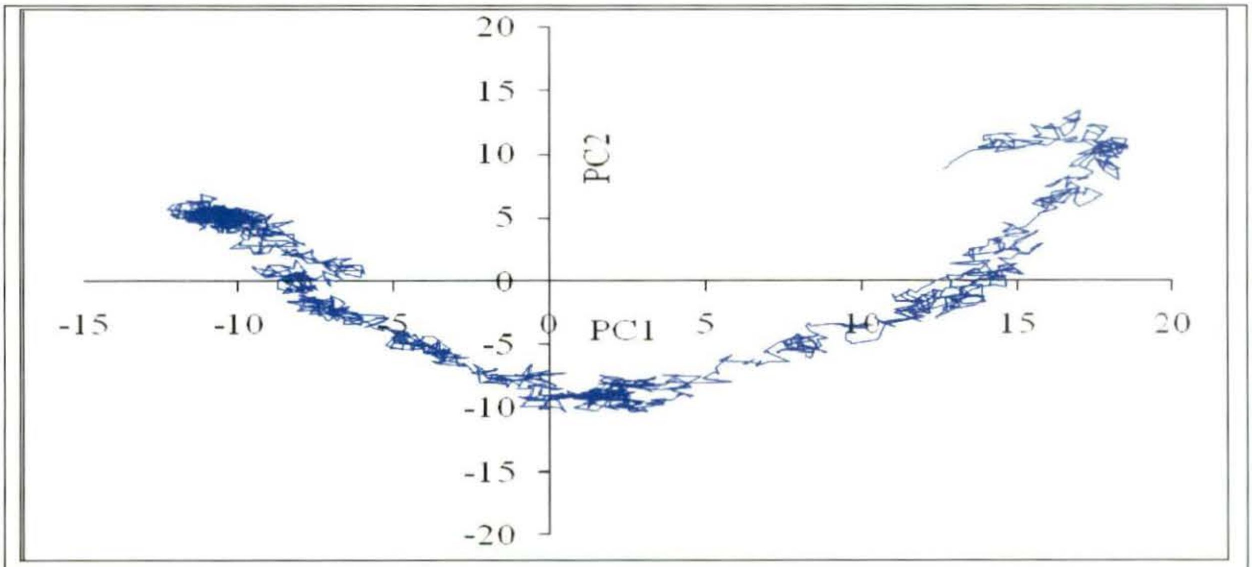


Figure-13: **Conformational Sampling**: The probability of sampling the phase space determined by principal modes 1 and 2 during the simulations of the modelled CdtB protein of *H. hepaticus* ATCC51449.

4. Discussions:

The modelled structures of the CdtB protein from *H. hepaticus* ATCC51449 shows that the helix and sheets remain interspersed throughout the protein structure (Fig. 2). VERIFY 3D revealed that 96.72% of the residues had an average 3D-1D score > 0.2.

Results of z scores from ProSA analysis specify that the z score of our model is very much within the range of scores normally found for proteins of comparable size. The value of the z-scores signifies that the 3D model of the CdtB protein is reliable and precise.

It is evident from Ramachandran plot (Fig. 5) that our predicted model protein CdtB carrying characteristics of a good quality of a model protein

Study of the molecular surface and electrostatic potential of modelled CdtB shows charged and polar residues are mostly on the surface. (Fig. 6).

It is evident from Fig.-7 that RMSD increased slowly up to 3,900 ps, and then decreases upto 4200ps then again slightly increases upto 5000ps. It is also clear from Rg, which shows a little variation upto 3000ps and after that almost attains equilibrium (Fig. 8). The RMSD and Rg calculations of CdtB suggest that the protein is less flexible in nature.

From RMSF, it is evident that the first and last residue fluctuates considerably. Interestingly, pronounced fluctuations are observed along some amino acid stretches (43-47, 75-81, 236-244,), which indicate the flexibility of the toxin in that region (Fig. 9).

The number of hydrogen bonds ranges from 133 to 196 (Fig. 10) revealed the fact that during the simulation, several hydrogen bonds broke and formed.

It is seen from Fig. 11 that time evolution of principal component 1 (PC1), principal component 2 (PC2), fluctuates remarkably in comparison to principal component 3 (PC3) and principal component 4 (PC4)

It is evident from the RMSF of Cα atoms calculated after projecting trajectories along their respective four principal components analysis for CdtB, fluctuation is highest in the projection on PC1 which indicates that PC1will provide more information regarding the collective motion of the protein (Fig. 12).

It is clear from the probability of sampling the phase space determined by first two principal modes during the simulations the projection of the dynamics trajectory onto the first two PC that the protein A traverse one conformational space around the origin and second one at the right side of the origin and third one at the left side of the origin which are much scattered indicating slightly high conformational freedom (Fig. 13)

The aim of our study was to construct three-dimensional model of the CdtB protein from *H. hepaticus* ATCC51449 using the homology modelling technique. The structures presented here are reliable on their biochemical features. The model may help to explain the functional mechanism of the Cdtb protein from *H. hepaticus* ATCC51449. The RMSD and Rg study explain the rigidity of toxin. This model is expected to assist the scientists working with the *H. hepaticus* ATCC51449 to recognize structure-function relationships of the CdtB protein. In absence of crystallographic or NMR structure this model will enlighten us about three dimensional structure and dynamic properties of the toxin CdtB and opening newer possibilities for exploring the molecular mechanism and activity in CdtB protein *H. hepaticus* ATCC51449.

8.5: References:

Altschul SF, Madden TL & Schaffer AA (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* **25**:3389-3402

Amadei A, Linssen ABM & Berendsen HJC (1993) Essential dynamics of proteins. *Proteins,* **17**:412–425

Avenaud P, Castroviejo M, Claret S, Rosenbaum J, Mégraud F & Ménard A (2004) Expression and activity of the cytolethal distending toxin of *Helicobacter hepaticus. Biochemical and Biophysical Research Communications* **318:**739–745

Berendsen HJC, Postma JPM, van Gusteren WF, Di Nola A & Haak JR (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81**:3684–3690

Bouzari S & Varghese A (1990) Cytolethal distending toxin (CLDT) production by enteropathogenic *Escherichia coli* (EPEC). *FEMS Microbiol. Lett.*, **59:**193–198

Centeno NB, Planas-Iglesias J & Oliva B (2005) Comparative modelling of protein structure and its impact on microbial cell factories. *Microbial Cell Factories.,* **4**:20 Doi: 10.1186/1475-2859-4-20

Cope L, Lumbley S, Latimer J, Klesney-Tait J, Stevens M, Johnson L, Purven M, Munson RJ, Lagergard T, Radolf J & Hansen E (1997) A diffusible cytotoxin of *Haemophilus ducreyi. Proc. Natl. Acad. Sci. USA.,* **94:**4056–4061

Darden T, York D & Pedersen L (1993) Particle mesh Ewald—an N.log (N) method for Ewald sums in large systems. *J. Chem. Phys.,* **98**:10089–10092

Das A & Mukhopadhyay C (2007) Application of principal component analysis in protein unfolding: An all-atom molecular dynamics simulation study. *The Journal of Chemical Physics,* **127**:165103-165108.

Dassanayake RP, Griep MA & Duhamel GE (2005) The cytolethal distending toxin B sub-unit of *Helicobacter hepaticus* is a Ca2+- and Mg2+-dependent neutral nuclease. *FEMS Microbiology Letters* **251**: 219–225.

Deng K & Hansen EJ (2003) A CdtA–CdtC complex can block killing of HeLa cells by *Haemophilus ducreyi* cytolethal distending toxin. *Infect. Immun.,* **71**: 6633–6640

Dlakic M (2000) Functionally unrelated signalling proteins contain a fold similar to Mg2þ-dependent endonucleases. *Trends Biochem. Sci.* **25**:272–273.

Dlakic M (2001) Is CdtB a nuclease or a phosphatase? *Science* **291**:547.

Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y & Liang J (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Research,* **34**:116-118.

Eisenberg D, Luthy R & Bowie JU (1997) VERIFY3D: Assessment of protein models with three-dimensional profiles. *Methods in Enzymology,* **277**:396-404.

Elwell C, Chao K Patelet K & Dreyfus L (2001) *Escherichia coli* CdtB mediates cytolethal distending toxin cell cycle arrest, *Infect. Immun.,* **69**:3418–3422.

Elwell CA & Dreyfus LA (2000) DNase I homologous residues in CdtB are critical for cytolethal distending toxin-mediated cell cycle arrest. *Mol. Microbiol.* **37:** 952–963.

Essmann U, Perera L, Berkowitz ML, Darden T, Lee H & Pedersen LG (1995) A smooth particle mesh Ewald method. *J. Chem. Phys.,* **103**:8577–8593

Frisan T, Cortes-Bratti X, Chaves-Olarte E, Stenerlöw B & Thelestam M (2003). The *Haemophilus ducreyi* cytolethal distending toxin induces DNA double strand breaks and promotes ATM-dependent activation of RhoA. *Cell. Micro.,* **5**:695-707

Garcia AE (1992) Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.,* **68**:2696–2699

Goodshell DS, Morris GM & Olson AJ (1996) Automated docking of flexible ligands: Applications of autodock, *Journal of Molecular Recognition*, 9:1-5

Guex N & Peitsch, MC (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*, **18**:2714-2723

Hassane, DC, Lee RB, Mendenhall MD & Pickett CL (2001) Cytolethal distending toxin demonstrates genotoxic activity in a yeast model. *Infect. Immun* , **69**:5752.

Hess B, Bekker H, Berendsen HJC & Fraaije JGE (1997) LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**:1463-1472.

Johnson WM & Lior H (1988) A new heat-labile cytolethal distending toxin (CLDT) produced by *Campylobacter* spp. *Microb. Pathog.,* **4**:115–126

Johnson WM & Lior H (1988b) A new heat-labile cytolethal distending toxin (CLDT) produced by *Escherichia coli* isolates from clinical material. *Microb. Pathog.,* **4**:103–113.

Lara-Tejero M & Galan JE (2002) Cytolethal distending toxin: limited damage as a strategy to modulate cellular functions. *Trends Microbiol.*, **10**:147–152.

Lara-Tejero M & Gala'n JE (2000) A bacterial toxin that controls cell cycle progression as a deoxyribonuclease I-like protein. *Science,* **290**: 354–357

Lara-Tejero M & Gala'n JE (2001) CdtA, CdtB, and CdtC form a tripartite complex that is required for cytolethal distending toxin activity. *Infect. Immun.*, **69**: 4358–4365

Laskowski RA, MacArthur MW, Moss DS & Thornton JM (1993) *PROCHECK*: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, **26**: 283-291

Laskowski RA, Watson JD & Thornton JM (2005) ProFunc: a server for predicting protein function from structure. *Nucleic Acids Research,* **33**:89-93.

Lee RB, Hassane DC, Cottle DL & Pickett CL (2003) Interactions of *Campylobacter jejuni* cytolethal distending toxin subunits CdtA and CdtC with HeLa cells. *Infect. Immun.,* **71**: 4883–4890

Lindahl E, Hess B & van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Modeling* **7**:306-317.

Mayer MPA, Bueno LC, Hansen EJ & DiRienzo JM (1999) Identification of a cytolethal distending toxin gene locus and features of a virulence-associated region in Actinobacillus actinomycetemcomitans. *Infect Immun.,* **67**:1227–1237.

Okuda J, Kurazono H & Takeda Y (1995). Distribution of the cytolethal distending toxin A gene (cdtA) among species of Shigella and Vibrio, and cloning and sequencing of the cdt gene from Shigella dysenteriae. *Microb. Pathog.,* **18**:167–172

Pickett CL & Whitehouse CA (1999) The cytolethal distending toxin family. *Trends. Microbiol.,* 7:292–297.

Pickett CL, Pesky EC, Cottle DL, Russell G, Erdem AN & Zeytin H (1996) Prevalence of cytolethal distending toxin production in *Campylobacter jejuni* and relatedness of Campylobacter sp. cdtB gene. *Infect. Immun.,* **64**:2070–2078

Rajesh R, Gunasekaran K, Muthukumaravel S, Balaraman K & Jambulingam P (2007) In Silico analysis of voltage-gated sodium channel in relation to DDT resistance in vector mosquitoes. *InSilico Biology,* 7:413-421.

Ramachandran GN, Ramakrishnan C & Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology,* 7:95-99.

Saiki K, Gomi T & Konishi K (2004) Deletion and purification studies to elucidate the structure of the *Actinobacillus actinomycetemcomitans* cytolethal distending toxin. *J. Biochem. (Tokyo),* **136**: 335–342

Saiki K, Konishi K, Gomi T, Nishihara T & Yoshikawa M (2001). Reconstitution and purification of cytolethal distending toxin of *Actinobacillus actinomycetemcomitans*. *Microbiol. Immunol.*, **45**: 497–506.

Sali A & Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology,* **234**:283-291.

Shenker BJ, Besack D, *McKay T, Pankoski L, Zekavat A & Donald R. Demuth DR* (2004) *Actinobacillus actinomycetemcomitans* cytolethal distending toxin (Cdt): Evidence that the holotoxin is composed of three subunits: CdtA, CdtB, and CdtC. *J. Immunol.*, **172:** 410–417

Suerbaum S, Josenhans C, Sterzenbach T, Drescher B, Brandt P, Bell M, Droge M, Fartmann B, Fischer HP, Ge Z, Hörster A, Holland *R,* Klein K, König J, Macko L, . Mendz GL, Nyakatura G, Schauer DB, Shen Z, Weber J, Frosch M & , Fox JG (2003) The complete genome sequence of the carcinogenic bacterium *Helicobacter hepaticus*. *PNAS.*, **100**:7901–7906

Sugai M, Kawamoto T, Peres SY, Ueno Y, Komatsuzawa H, Fujiwara T, Kurihara H, Suginaka H & Oswald E (1998) The cell cycle-specific growth-inhibitory factor produced by *Actinobacillus actinomycetemcomitans* is a cytolethal distending toxin. *Infect. Immun.*, **66:**5008–5019

Thompson JD, Higgins DG & Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research,* **22**:4673-4680.

Ward JM, Fox JG, Anver MR, Haines DC, George CV Collins Jr. MJ, Gorelick PL, Nagashima K, gonad MA, gilden RV, Tully JG, Russell R, Benvensite RE, Paster BJ, Dewhirst, FE, Donovan JC, Anderson LM & rice JM (1994) Chronic active hepatitis and associated liver tumors in mice caused by a persistent bacterial infection with a novel *Helicobacter* species. *J. Natl. Cancer Inst.,* **86**:1222–1227

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

Yamada T, Komoto J, Saiki K, Konishi K & Takusagawa F (2006) Variation of loop sequence alters stability of cytolethal distending toxin (CDT): Crystal structure of CDT from *Actinobacillus actinomycetemcomitans. Protein Science.,* **15**:362–372

Yang LW, Rader AJ, Liu X, Jursa CJ, Chen SC, Karimi HA & Bahar I (2006) oGNM: A protein dynamics online calculation engine using the Gaussian Network Model. NucleicAcids Res., **34**:W24–31.

Young VB, Knox KA & Schaue DB (2000) Cytolethal distending toxin sequence and activity in the enterohepatic pathogen *Helicobacter hepaticus. Infect. Immun.,* **68**: 184–191.

APPENDIX

Appendix I: Software used in molecular dynamics simulation.

1. Abalone (classical, implicit water)
2. ABINIT (DFT)
3. ACEMD (running on NVIDIA GPUs: heavily optimized with CUDA)
4. ADUN (classical, P2P database for simulations)
5. AMBER (classical)
6. Ascalaph (classical, GPU accelerated)
7. CASTEP (DFT)
8. CPMD (DFT)
9. CP2K (DFT)
10. CHARMM (classical, the pioneer in MD simulation, extensive analysis tools)
11. COSMOS (classical and hybrid QM/MM, quantum-mechanical atomic charges with BPT)
12. Desmond (classical, parallelization with up to thousands of CPU's)
13. Culgi (classical, OPLS-AA, Dreiding, Nerd, and TraPPE-UA force fields)
14. DL_POLY (classical)
15. ESPResSo (classical, coarse-grained, parallel, extensible)
16. Fireball (tight-binding DFT)
17. GROMACS (classical)
18. GROMOS (classical)
19. GULP (classical)
20. Hippo (classical)
21. Kalypso MD simulation of atomic collisions in solids
22. LAMMPS (classical, large-scale with spatial-decomposition of simulation domain for parallelism)
23. LPMD Las Palmeras Molecular Dynamics: flexible an modular MD.
24. MacroModel (classical)
25. MDynaMix (classical, parallel)
26. MOLDY (classical, parallel) latest release
27. Materials Studio (Forcite MD using COMPASS, Dreiding, Universal, cvff and pcff forcefields in serial or parallel, QMERA (QM+MD), ONESTEP (DFT), etc.)
28. MOSCITO (classical)
29. NAMD (classical, parallelization with up to thousands of CPU's)
30. nano-Material Simulation Toolkit
31. NEWTON-X (ab initio, surface-hopping dynamics)
32. ORAC (classical)
33. ProtoMol (classical, extensible, includes multigrid electrostatics)
34. PWscf (DFT)
35. RedMD (coarse-grained simulations package on GNU licence)
36. S/PHI/nX (DFT)
37. SIESTA (DFT)
38. VASP (DFT)
39. TINKER (classical)
40. YASARA (classical)
41. XMD (classical)

Related software

1. VMD- MD simulation trajectories can be visualized and analyzed.
2. PyMol - Molecular Visualization software written in python
3. Packmol Package for building starting configurations for MD in an automated fashion
4. Sirius - Molecular modeling, analysis and visualization of MD trajectories
5. esra - Lightweight molecular modeling and analysis library (Java/Jython/Mathematica).
6. Molecular Workbench - Interactive molecular dynamics simulations on your desktop
7. BOSS - MC in OPLS
8. Punto is a freely available visualisation tool for particle simulations.

Specialized hardware for MD simulations

1. Anton - A specialized, massively parallel supercomputer designed to execute MD simulations.
2. MDGRAPE - A special purpose system built for molecular dynamics simulations, especially protein structure prediction.

Appendix II : Software used in Homology modeling

Name	**Method**	Description
LOMETS	Local Meta threading server	Meta-server combining 9 different programs
3D-JIGSAW	Fragment assembly	Automated webserver
Biskit	wraps external programs into automated workflow	BLAST search, T-Coffee alignment, and MODELLER construction
CABS	Reduced modeling tool	Downloadable program
CPHModel	Fragment assembly	Automated webserver
ESyPred3D	Template detection, alignment, 3D modeling	Automated webserver
FoldX	Energy calculations and protein design	Downloadable program
GeneSilico	Consensus template search/fragment assembly	Webserver
Geno3D	Satisfaction of spatial restraints	Automated webserver
HHpred	Template detection, alignment, 3D modeling	Interactive webserver with help facility
LIBRA I	LIght Balance for Remote Analogous proteins, ver. I	Webserver
MODELLER	Satisfaction of spatial restraints	Standalone program in Python
EasyModeller	GUI to MODELLER	Standalone windows executable
Protinfo CM	Comparative modelling of protein structure using minimum perturbation and looop building	Web server
ROBETTA	Rosetta homology modeling and ab initio fragment assembly with Ginzu domain prediction	Webserver
Selvita Protein Modeling Platform	Package of tools for protein modeling	Free demo, interactive webserver and standalone program including: BLAST search, CABS modeling, 3D threading, Psi-Pred secondary structure prediction
SWISS-MODEL	Local similarity/fragment assembly	Automated webserver based on ProModII