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 staphylococcal 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:

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

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:

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

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):

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

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:

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

where, ω_k signifies the relative adaptedness of the k^{th} 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:

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

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,

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

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

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

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

$s_l = -g_l$

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

$S_k = -g_k + b_k s(k-l)$

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:

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

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

$$a^p(t+\partial t) = a(t) + b(t)\partial t + \ldots\ldots \qquad (10)$$

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-∂t), v(t-2∂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+∂t and hence the correct accelerations a^c (t+∂t). These can be compared with the predicted acceleration from equation (10)

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

$$\Delta a(t+\partial t) = a^C(t+\partial t) - a^P(t+\partial t) \qquad (11)$$

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

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

The idea is that r^c (t+∂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+∂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 $t-\partial t$ to calculate new positions at time $t+\partial t$. 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 Virulence 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