2. Review of Literature

2.1 Discovery of nitrogen fixation

Early civilizations like that of the Egyptians knew the importance of legumes for maintaining soil fertility. The significance of crop rotation, green manuring and intercropping systems was practiced by major civilizations like the Romans and the Greeks. However, it was only in the 19th century that the role of microorganisms in nitrogen accumulation and thus crop improvement came to be recognized. JB Lawes and JH Gilbert, two British scientists in the late 19th century worked on the diverse capacities of legumes and cereals to accumulate nitrogen but failed to understand the mechanism (Lawes and Gilbert, 1895). However, the onus of the discovery fell upon two German scientists, Hellriegel and Wilfarth (1888) who identified that root nodules of leguminous plants were responsible for conversion of atmospheric nitrogen to ammonia. Martinus Beijerinck a Dutch scientist, isolated and cultured the microorganisms from root nodules of leguminous plants in 1888 (Sen, 1996). Beijerinck primarily named it *Bacillus radicicola* and proved that it helps in formation of nodules in roots of leguminous plants (SubbaRao, 1988). The modifications in the culture media took place over time and the nodulating bacterium was easily isolated and identified to be *Rhizobium*. The culture and isolation of nodulating bacteria from leguminous plants by Beijerinck opened the possibility of characterizing similar organisms in different parts of the world. Besides, a number of free-living nitrogen fixing microorganisms like *Clostridium pasteurianum* (Winogradsky, 1893), *Azotobacter chroococcum, Beijerinckia* (SubbaRao, 1988), *Klebsiella pneumoniae*, *Azospirillium* etc. were discovered.

2.2 Legume and non legume symbiosis

Some microorganisms enter into

symbiotic association with a number of leguminous as well as non-leguminous plants. Symbioses with higher plants offer an ecological niche for the particular microbe to fix nitrogen (Mylona *et al.,* 1995). The rhizobia groups of bacteria consisting of the genera like *Rhizobium, Bradyrhizobium, Sinorhizobium, Azorhizobium* and *Mesorhizobium* enter into symbiosis with the members of the legume family (Mylona *et al.,* 1995). Different species of *Rhizobium* generally forms association with peas, lucerne, clover and beans while, *Bradyrhizobium* and *Azorhizobium* forms associations with soyabeans and *Sesbania* respectively (Schlegel, 1991). Most of the legume symbionts are generally host specific (Mylona *et al.,* 1995).

Since the discovery of nitrogen fixation, legume-rhizobia association has been the most studied relationship owing to its importance in improving soil fertility (SubbaRao, 1988). Symbiotic association is not limited to the legumes but to a number of non-legumes. The most significant amongst them are the actinorhizal plants-*Frankia* association. This association involves a number of plants like *Casuarina, Hippophae, Alnus, Myrica,* etc. belonging to different families (Benson and Silvestor, 1993). Some cereals like rice, wheat, maize, millets, grasses etc. have symbiotic association with microorganisms accomplishing the process of nitrogen fixation (Saikia and Jain, 2007). Prominent among these microbes are *Azospirillium, Acetobacter* and *Azoarcus* (Saikia and Jain, 2007). Besides, the *Anabaena-Azolla* association (Bohlool *et al.,* 1992), *Nostoc–Gunnera* association (Mylona *et al.,* 1995) can fix a substantial amount of nitrogen. Some workers (Willey, 1979) suggest that in semi-arid tropics legumes play a leading role compared to humid tropics and intercropping system has been practiced with cereals (Fujita *et al.,* 1992). In temperate climates, soybeans (Keyser and Li, 1992) and forages (Ledgard and Steele, 1992) are widely used. Isolation of microorganisms from surface-sterilized roots led to the discovery of nitrogen fixing endophytes (Döbereiner 1992; Baldani and Baldani, 2005). Cycads in association with cyanobacterial species can fix nitrogen (Costa and Lindblad, 2002).

In the rhizobia and *Frankia* the root nodules are the sites for nitrogen fixation. An important feature of the root nodules in leguminous plants is the existence of a unique red pigment called

leghaemoglobin. It helps in the oxygen uptake by terminal oxidases and boosts ATP production for nitrogenase activity (Schlegel, 1991). Nodule formation is not only limited to the roots but in stems of *Aeschynomeme* and *Sesbania* and leaves of *Ardisia*, *Pavetta* and *Psychotria* (SubbaRao, 1988). However, bacteria associated with leaf nodules do not fix nitrogen. Compared to legume symbiotic microorganisms the first isolation of *Frankia* a non legume symbiont strain was achieved from nodules of *Comptonia peregrina* in 1978 (Callaham *et al.*, 1978). *Frankia* has two distinctive developmental structures, vesicles and spores crucial for its survival (Lechevalier, 1994). Vesicles are the sites for nitrogen fixation, while spores contained in are the reproductive structures in *Frankia.* In symbiotic nitrogen fixation, the establishment of the bacterium inside the host root and nodule development is a complex process and involves the processes like recognition and infection of the host root, differentiation of the root nodules, proliferation of the bacteria and transformation of the bacteroids in nodules (Schlegel, 1991).

2.3 Free-living nitrogen fixers

Free living microorganisms pass independent life and fix nitrogen. Aerobic free living microbes include *Azotobacter, Beijerinckia, Azonomas* etc., facultative anerobes include *Klebsiella*, anaerobic bacteria like *Clostridium acetobutylicum*, *C. pasteurianum*, photosynthetic bacteria like *Chlorobium*, *Rhodospirillium* etc. (Schlegel, 1991). It has been established (SubbaRao, 1988) that among cyanobacteria both heterocystous and non heterocystous forms, *Anabaena*, *Nostoc*, *Trichodesmium*, *Lyngbya*, *Plectonema* etc. fix nitrogen (Schlegel, 1991). These occur in wide ranging habitats and ecological niches. Methanogenic Archaea gave a wider angle to the field of biological nitrogen fixation. Zinder and Daniels were the first to discover nitrogen fixation in different methanogenic Archaea (Murray and Zinder, 1984; Belay *et al.*, 1984). In cyanobacteria a light and thick walled cell structure called heterocyst is the point where nitrogen fixation takes place. However, in members like *Lyngbya*, *Plectonema* etc. where heterocyst is absent nitrogen fixation occur in internally organized cells (Schlegel, 1991).

2.4 Nitrogenase

The most important enzyme associated with the nitrogen fixing mechanism is nitrogenase. It is a quite complex

biological molecule that reduces dinitrogen to ammonia at appropriate temperature and pressure (Peters and Szilagyi, 2006). SubbaRao (1988) reported that Carnahan and his co-workers in the year 1960 were the pioneers in carrying out the reduction of dinitrogen to ammonia in cell free extracts of *Clostridium pasteurianum*. This discovery ushered the concept of universality of nitrogenase in free-living as well as symbiotic nitrogen fixing systems. Mo-nitrogenase is the conventional nitrogenase, harboring a prosthetic group with molybdenum. *Azotobacter* and some photosynthetic dizaotrophs, bear extra type of nitrogenase whose cofactor houses vanadium or only iron (Newton, 2007). The conventional nitrogenase is an equilibrium mixture of molybdo-ferro protein and iron protein in the ratio of 1:2. Two FeMoCo are attached to α subunits of the MoFe protein. Besides, 4Fe- 4S clusters comprise of some other prosthetic groups. The P-clusters are covalently linked to cysteine residues of molybdo-ferro protein forming a bridge between α and ß subunits (Rubio and Ludden, 2008). Alexander (1977) stated that all nitrogen fixers had alike proteins, signifying that a similar but not matching reaction sequence. Consequently the molecular weights in the two components differ (Johanson *et al.*, 1983). In the later part of the 20th century studies on the enzymology and structures of nitrogenase became appropriate since, nitrogenase catalysis reactions involved biochemical processes like protein-protein interactions, signal transduction and electron-transfer reactions. The studies concerning the enzymology of nitrogenases were hampered due to the dearth of the establishment of a complete mechanism for reduction of dinitrogen with respect to the structure and properties of the nitrogenase proteins (Howard and Rees, 1996). The establishment of the three-dimensional structures for the nitrogenase proteins in 1992 and their metal clusters considerably changed the research on nitrogen fixation (Georgiadis *et al.*, 1992; Kim *et al.*, 1993; Einsle *et al.*, 2002). The establishment of the structure paved the way for developments in the areas of biochemistry, spectroscopy, biophysics, protein chemistry in describing structural features of the nitrogenase proteins, their functional mechanism and capacity to catalyse nitrogen fixation under ambient conditions by various workers (Burgess and Lowe,

1996; Rees and Howard, 2000; Lawson and Smith, 2002; Seefeldt *et al.*, 2004).

2.5 Genetics of nitrogen fixation

A wide array of techniques like mutations, deletion mapping, cloning vectors etc. have facilitated the identification of genes associated with nitrogen fixation. Legume, non legume and free living nitrogen fixers have a set of genes which are responsible for effective nodulation and nitrogen fixation. These are the *nod, nol, noe, nif, fix* and some hydrogenase genes. The work on the genetics of nitrogen fixation was first started in *Klebsiella oxytoca* M5a1 and first ever detailed organization of *nif* genes were reported in this organism (Arnold, 1988). A number of studies (Zheng *et al.*, 1998; Merrick and Edwards, 1995; Dixon and Kahn, 2004; Hu *et al.*, 2007; Rubio and Ludden, 2008) have established that core *nif* genes like *nifH, nifD nifK, nifY, nifB, nifQ, nifE, nifN, nifX, nifU, nifS, nifV, nifW, nifZ* are essential for nitrogen fixation. On the basis of mutational studies the natures of different *nif* gene products were determined. Studies confirmed that *nif* HDK encodes nitrogenase while *nif* LA had regulatory functions. In cyanobacteria especially *Anabaena* sp. 7120 it was established that the organization of *nif*HDK differed from that of *Klebsiella pneumoniae* (Godlen *et al.*, 1987). Johnston and his co-workers first discovered the presence of nodulation genes in a plasmid of *Rhizobium leguminosarum* (SubbaRao, 1988) and mutation of those genes rendered them useless. Later on studies (Schukze and Kondorosi, 1998; Perret *et al.*, 2000) ascertained that *nod, nol* and *noe* genes produce nodulation signals. The interplay of different *nod* genes, triggering of the creation of root nodule, signaling cascades and development of nodule meristem were reported by a number of researchers. (Yang *et al.*, 1999; Long, 2001; Geurts and Bisseling, 2002). Sprent and James (2007) reviewed the evolution of different nodule structures and infection types in leguminous plants. Studies concerning nodule physiology and nodule functioning specified that the bacteria in bacteroids forms are surrounded in plant membranes (White *et al.*, 2007) resulting in the formation of symbiosomes. Ammonia produced in bacteroids is transported to the plant all the way through the symbiosomes by amino acid flux system (White *et al.*, 2007). Besides, alanine is transported. The origin of nodulation genes in leguminous plants are clouded in

mystery however research is on in this field. Like the rhizobia, *Azosprillium* includes a megaplasmid and sequences similar to *nod* genes (Elmerich, 1987). *Frankia* houses a number of *nif* genes but researchers failed to spot *nod* genes in *Frankia* (Cérémonie *et al.*, 1998). However, groundwork on *Frankia* genomes exposed some putative *nod*-like genes which did not emerge in organized clusters and failed to detect the *nod* A gene (Franche *et al.*, 2009). SubbaRao (1988) reported that *fix* genes were identified on the chromosomes as well as plasmids of *Rhizobium* and were transferable. These genes are irregular in free-living bacteria (Gubler and Hennecke, 1986) and assist in electron transfer to the nitrogenase.

2.6 Pre-genomic era research in biological nitrogen fixation (BNF)

The discovery of nitrogen fixing organisms throughout the globe, studies concerning their gene products as well as the development of molecular biology helped BNF research enter a new stage where basic research combined with latest techniques. BNF research in the late 20th century mainly focused on exploration and application of diazotrophs resources, mechanisms of ammonium suppression, oxygen sensitivity of *nif* gene expression, regulation between plant and microorganisms during symbiosis, nature and synthesis of nodulation factor, biochemistry and functions of nitrogenase and increase in nitrogen fixation efficiency for crop productivity (Franche *et al.*, 2009). Advances in gene cloning, PCR amplifications and sequencing of DNA ushered a new way of identifying diazotrophs. Zehr and McReynolds (1989) were the pioneers in developing a number of primers for amplification of *nif*H fragments from the environment. It enabled the estimation of the diversity of diazotrophs in the rhizosphere without going through the complex process of bacterial isolation, cultural studies biochemical studies etc. (Hamelin *et al.*, 2002; Roesch *et al.*, 2008). Stoffels *et al.*, (2001) reported that studies on 16S rRNA genes design of specific probes in the later part of the 20th century.

2.7 Beginning of interdisciplinary research and dawn of bioinformatics

It was in the year 1986 that the Department of Energy (DOE), USA and National Institute 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). According to Kyrpides (1999) 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. 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 (2001) 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 20^{th} 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. On the other side development of computers and sophisticated operating systems were also going on. The publication of huge amount of sequence data were greatly supported by

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. On one hand the science of bioinformatics developed and on the other hand 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.

2.8 Post genomic era and BNF

The amalgamation of the knowledge of plant physiology, biochemistry, genetics and molecular biology gave idea about the understanding of the mechanism of nitrogen fixation in pre-genomic era. Nevertheless, the knowledge remained erratic and discrete (Benedito *et al.*, 2006). Things changed with the accessibility of complete genome sequences of symbiotic as well as non-symbiotic diazotrophs. Knowledge of the whole genome became the stepping stone in understanding the working principle of the bacterial cell (Puhler *et al.*, 2004). *Mesorhizobium loti* was the first sequence of a symbiotic bacterium and it was followed by *Sinorhizobium meliloti* (Puhler *et al.*, 2004). The completion of the genomes of *Rhizobium leguminosarum bv viciae* (Young *et al.*, 2006), *Rhizobium etli* (Gonza´lez *et al.*, 2006), *Bradyrhizobium* strains (Kaneko *et al.*, 2002), *Frankia* strains (Normand *et al.*, 2007) and sequences for a number of free-living diazotrophs spanning different habitats and ecological niches (Franche *et al.*, 2009) bolstered nitrogen fixation research. The studies on the genomes exposed new evidences pertaining to evolution and structure, interactions between plants and microbes and diversity amongst the diazotrophs. The breakthrough in unearthing the gene maneuvering propagation of nodule primordia opened a new dimension in functional genomics of rhizobia-legume symbiosis (Gresshoff, 2003). The detection of a number of symbiotic genes associated with nitrogen fixation has strengthened functional genomic research. Tools like

DNA macro- and microarrays have been applied for studying expression at transcription level in *Sinorhizobium* and are being applied in other rhizobia (Puhler *et al.,* 2004). Genome sequences of *Frankia* yielded significant information regarding their evolutionary histories and linked the disparity in size with the biogeographic history of the host plants of each strains (Normand *et al.,* 2007). Transcriptomic and proteomic investigations of the sequenced genomes provided important information regarding rhizobial-legume symbioses (Puhler *et al.,* 2004). The analysis of the genomes of *Azoarcus* and *Gluconacetobacter* is expected to start comparative genomic studies on endophytic as well as endosymbiotic associations (Puhler *et al.,* 2004). MacLean *et al.,* (2007) reported that genome studies have proved that the architecture of the genome content is guided by the lifestyle of legume symbionts. Since, legumes are fundamental part of agriculture and a significant component of sustainable agriculture, sequencing projects of legumes viz. *Lotus japonicus*, *Glycine max*, *Medicago trunculata* got underway (Udvardi, 2002) for better understanding of the symbiotic interactions. Sequencing projects on these legumes focused on the large number of EST's (Expressed sequence tags). According to Udvardi (2002) until 2002, the public database GenBank housed 255,291, 162,741 and 31,670 ESTs for these species. He further stated it has become a gold mine since EST collections symbolized cDNA libraries resulting either from specific organs obtained in different developmental stages, or from plants subjected to biotic and abiotic stress conditions and can be used for studying expression levels. The protein patterns in the diazotrophs are being studied using advanced proteomics approaches. Proteomic analyses have also revealed the direct genome functionality in a number of diazotrophs genomes (MacLean *et al.,* 2007). The understanding of the expression level of the genes is going on and is expected to throw light on the adaptation of the diazotrophs in soil or in symbiotic association. The completion of the genome projects for a large number of marine cyanobacteria as well as free-living diazotrophs like *Chlorobium, Clostridium, Azotobacter* and some methanogenic archaea has revealed new surprises and functional genomics approaches are being applied in a large scale. Studies concerning the nitrogen fixation in open oceans offer

wealth of information about marine diazotrophs. There has been a revolution in the field of ocean genomics. The sequencing projects for *Synechococcus, Trichodesmium* etc., discovery of novel genes yielded considerable information regarding their genome architecture in their respective habitats (Zehr *et al.*, 2000).

The research on functional genomics and proteomics for the nitrogen fixing 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 diazotrophs.

2.9 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.*, 1980b; Grantham *et al.*, 1981). 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 conjectured 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 and

Gautier, 1982) in *E.coli* regarding codon usage and tRNA abundance led to the inference that highly expressed genes exhibited non-random codon usage and used a miniature set of codons that corresponded to abundant tRNAs. However, it was not apparent why specific synonymous codons were used preferably. Grosjean and 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) first hypothesized 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. Some workers (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 and Mrazek, 2000). Majority of the highly expressed genes are associated with cell growth and cell division. It has been found that evaluating translational selection is 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 and Yang, 1994; Nesti *et al.*, 1995; Pouwels and Leunissen, 1994; Schmidt, 1995) helps in determining the phylogenetic relationships between

organisms and inferring significant evolutionary trends. 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 and Gautier, 1994) in a prominent manner. 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. 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 (Peden, 1999). 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 and Gautier, 1982) measuring the codon-anticodon interaction, codon bias index (CBI) (Bennetzen and 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 and 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 and 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 prokaryotes work on nitrogen fixing microorganisms has been lacking. Very preliminary work on codon usage of nitrogen fixing diazotrophs was initiated by Mather and Tuli (1991). McEwan and Gatherer (1999) used certain indices to predict gene functionality in an operon of *Frankia.* Ramseier and Gottfert (1991) reported differences in codon usage and GC content in *Bradyrhizobium* genes. Musto *et al.,* (2003) accounted the role of translational selection for synonymous codon usage in *Clostridium acetobutylicum,* however they did not deal with the nitrogen fixation related genes. All these are just very little and a comprehensive work with focus on lifestyle patterns are the need of the hour.

2.10 Proteomics and structural bioinformatics of nitrogen fixation

High throughput sequencing of a number of eukaryotes and a number of bacteria has developed the science of

proteomics and 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 and Bonnano, 2002). 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 nitrogen fixation research.

Proteomics techniques have been used to explore legume-microbe interactions. The proteomic studies are focusing on determination of the influence of microorganisms on host roots, for identifying genes and proteins induced during root-microbe interactions (Mathesius, 2009). Most of the research has dealt with legume-symbioses for nitrogen-fixing rhizobia. It has been possible to detect specific microbial signal molecules in plants, harmonizing of signal responses and modification of plant growth by microorganisms. Combinations of 2D gel, peptide mass fingerprinting and bioinformatics aided in understanding of factors linked to interactions of *Rhizobium*, *Mesorhizobium*, *Sinorhizobium* with model legume plants like *Medicago trunculata* and *Lotus japonicus* (Rolfe *et al.*, 2003). Investigation concerning acid tolerance of *Bradyrhizobium japonicum* USDA 110 using 2-D gel electrophoresis and MALDI-TOF techniques offered information regarding protein compositions under variable pH and illustrated that acid tolerant property of *Bradyrhizobium japonicum* had relationships with a number of important metabolic pathways (Puranamaneewiwat *et al.*, 2006). A number of proteomic approaches were used to identify cyanobacterial symbiotic associations in *Azolla*. The analysis revealed some metabolisms that were committed to nitrogen fixation and expression of some vital proteins earlier unknown to influence cyanobacterial symbiosis. Further analysis exposed the nature of modifications associated with NifH proteins in the cyanobacteria (Ekman *et al.*, 2008). Proteomics studies of *Gluconacetobacter diazotrophicus* in exponential and stationary phases of cultures in presence of low and high levels of combined nitrogen revealed 131 differentially expressed proteins (Lery *et al.*, 2008). Proteins associated with cell wall biogenesis were reported

(Lery *et al.*, 2008) to be expressed under limited nitrogen and shaped functioning of nitrogen-fixing cells. Nitrogen fixation accessory proteins were produced when cells grew under limited nitrogen and these proteins helped in stabilization of nitrogenase as well its protection from oxygen. Comparative proteomic analysis of *Rhodospirillum rubrum* identified 44 differentially expressed proteins. Proteins stimulated in nitrogen-fixing conditions are associated with nitrogen fixation and assimilation and are consistent with Nif regulated expressions (Selao *et al.*, 2008).

Although a number of works has been performed using X-ray crystallography and NMR spectroscopy for *Azotobacter* and *Clostridium* nitrogenases (Burgess and Lowe 1996; Rees and Howard 2000; Lawson and Smith 2002; Seefeldt *et al.*, 2004) there are absence of comprehensive studies using homology modeling techniques. This might have been due to difficulties associated with dearth of appropriate crystal structures in the public domain that can be used as suitable templates coupled with non-availability of appropriate software and tools for validations. The difficulties in ascertaining the three-dimensional structures of proteins by spectroscopy and X-ray crystallographic analysis have made computational techniques very popular (Othman *et al.*, 2007). Spectroscopic and X-ray crystallographic based methodologies are robust, time-consuming, costly and often yield unreliable results. Homology modelling has become a dependable procedure for consistently predicting 3D structure of proteins with accuracy levels akin to those attained at low-resolution using experimental methods (Martin-Renom *et al.*, 2000). In this method the amino acid sequence of unknown structure (target) is aligned with the homologue of known structure (template). In view of the fact, that study of 3D structure of a protein is supportive in making out the finer points of a protein (Paramsivasan *et al.* 2006), this method has become wide spread in bioinformatics. Homology models of NifH proteins from *Trichodesmium* sp. IMS101 have reported by Zehr *et al.*, (1997) and the implications of aerobic nitrogen fixation were analyzed. Sen and Peters (2005) reported the homology model of the *nif*H protein from *Methanobacter thermoautotrophicus*. The structure demonstrated the interplay of a number of factors playing a role in the thermal adaptation of protein-protein

interactions in *M. thermoautotrophicus*. The structure has been reported to be rigid and assisted in thermal stability.

Most of the studies on the protein complement of the sequenced diazotrophs have largely focused on functional annotations and amino acid compositions with little focus on common physical properties. The steady increase in complete proteomes for diazotrophs pose a significant challenge for accomplishing study in large numbers. Thus, a broad picture on protein signatures for nitrogen fixing microorganisms is unavailable. The physical properties of proteins are vital for functioning of an organism (Knight *et al.,* 2004). An understanding of how these properties render the role of complete proteomes is expected to fetch answers relevant to organism and molecular biology (Knight *et al.,* 2004).

2.11 Phylogenetic studies on BNF

Nothing in biology makes sense without the light of evolution. This also applies for biological nitrogen fixation. Earlier studies of biological evolution mainly concentrated either upon general appearance or habitats or genetic similarities (Williams, 2007). The foundation and current distribution of nitrogen fixation has been mystifying from a phylogenetic point of view owing to a number of factors that puzzled the scientists for decades (Raymond *et al.,* 2004). Raymond *et al.,* (2004) reported that the ancient phenomenon of nitrogen fixation played a vital role in the spreading out of microbial life on our planet. All nitrogen fixing diazotrophs have a wide distribution in the kingdoms of Bacteria and Archaea. A number of workers (Raven and Yin, 1998; Kasting and Siefert, 2001; Navarro-Gonzalez *et al.,* 2001) hypothesized that abiotic nitrogen was a limiting factor in the early days of the Earth and reduced availability of nitrogen influenced the evolution of early nitrogen fixers. The complex climatic, geological and biochemical events of the early atmosphere too played a significant role in the evolution of nitrogen fixers. In the post genomic era, the accessibility of the genes, proteins and completely sequenced genomes have made it feasible to comprehend the impact of complex genetic events like paralogy, lateral gene transfer, gene duplication events in the evolution of nitrogen fixation. It has now become possible to understand clearly the forces influencing the evolution of nitrogenase at the molecular level. Raymond *et al.,* (2004)

reported that nitrogenase evolved in multiple lineages and there are evidences of loss, duplications and horizontal and vertical transfers for the nitrogenase genes and operons during the course of evolution. Till date most of the studies concerning the evolution of nitrogen fixation have focused on the *nif* genes since they are extensive, found in cotranscribed regulons, house the major subunits for the nitrogenase and code for a number of proteins involved in the regulation and activation of the nitrogen fixing mechanism (Kessler *et al.,* 1998; Kessler and Leigh, 1999; Halbleib and Ludden, 2000; Goodman and Weisz, 2002). Owing to the far-reaching efforts of scientists throughout the world it has been possible to sequence the nitrogenase genes like *nif*H, *nif*D, *nif*K, *nif*E, *nif*N etc. for understanding the evolutionary history of nitrogen fixation (Normand and Bousquet, 1989; Normand *et al.,* 1992; Hirsch *et al.,* 1995; Zehr *et al.,* 1997; Fani *et al.,* 2000). There has been a rigorous debate on the evolution of these genes among different workers (Normand and Bousquet, 1989; Fani *et al.,* 2000; Zehr *et al.,* 2003; Henson *et al.,* 2004; Kechris *et al.,* 2006). These debates focused on issues ranging from random distribution of nitrogen fixation (Henson *et al.,* 2004) to lateral gene transfer (Normand and Bousquet, 1989; Kechris *et al.,* 2006), vertical descent accompanied by multiple losses (Young, 1992; Normand *et al.,* 1992; Hirsch *et al.,* 1995), convergent evolution (Postgate and Eady, 1988) etc. There have been conflicting reports on the evolution of the genes arising out of investigations with dissimilar techniques of phylogenetic analysis (Henson *et al.,* 2004). Some workers (Normand and Bousquet, 1989; Normand *et al.,* 1992; Hirsch *et al.,* 1995; Hennecke *et al.,* 1992; Zehr *et al.,* 1998; Zehr *et al.,* 2000) reported contradictory information on the agreement between *nif*H phylogenies and 16S rRNA based phylogenies in cyanobacteria, proteobacteria and actinobacteria while others (Henson *et al.,* 2004) investigated the evolutionary history of nitrogen fixation using only *nif*D. Most of the workers used the sequence alignment based methods to study evolution of relevant *nif* genes.

Normand and Bousquet (1989) were the first to hypothesize that nitrogen fixation arose once in the course of evolution and has been transmitted horizontally to different lineages. The other hypothesis that cropped up in the last decade of the 20th century, predicted

that nitrogen fixation arose in the early evolutionary history and was omnipresent in all prokaryotic organisms at that time. Gradually with time it disappeared in some lineages and presently retained in some (Young, 1992; Normand *et al.,* 1992; Hirsch *et al.,* 1995). Fani *et al.,* (2000) postulated a model that in some ways recreated the evolutionary history of *nif*D, *nif*K, *nif*E and *nif*N. The primitive ancestral gene arose by means of paralogous duplication events (Fani *et al.,* 2000). This was followed by deviation to an ancestral bicistronic operon and this operon further went through paralogous operon duplication events thus creating the modern ancestors of *nif* genes. They further hypothesized that *nif DKEN* genes were already present in the earliest prokaryotes prior to the division of archaea and bacteria and weakening of selection pressure owing to higher energetic costs of the process of nitrogen fixation resulted in loss of *nif* genes in the late prokaryotes thus limiting them to a few. Henson *et al.,* (2004) re-examined the phylogeny of nitrogen fixation by analyzing only the molybdenum containing *nif*D gene from a cyanobacteria, proteobacteria as well as gram-positive bacteria. They compared the *nif*D phylogeny with that of the 16S rRNA phylogeny. The rationale behind this was that in cases where *nif* and rRNA topologies matched, vertical descent could be inferred otherwise lateral gene transfer could be the reason. They applied the parsimony technique, maximum likelihood technique and neighbor joining technique to analyze the evolution of the *nif*D genes. Henson *et al.,* (2004) inferred that their studies on *nif*D phylogenies are harmonious with that of the 16S rRNA based phylogenies and supported the hypothesis of vertical descent. Their results also matched with the *nif*H phylogenies (Zehr *et al.,* 1998, 2000) with respect to vertical descent. The parsimony analysis of the amino acid sequences and maximum likelihood studies concerning the nucleotide sequences of *nif*D genes support monophyletic origins of cyanobacterial and actinobacterial strains strengthening their hypothesis on vertical descent (Henson *et al.,* 2004). Conversely, distance matrix analysis of nucleotide sequences for *nif*D genes bolstered the concept of horizontal gene transfer. They also argued that the parsimony and maximum likelihood methodologies are probably more superior compared to other methodologies based on sequence alignments for resolving the