1. Introduction

Nitrogen is a stable, inert element comprising 78% of the Earth's atmosphere. Inspite of its abundance a small percentage of it is capable of entering into the biogeochemical cycle. Approximately, 0.001 per cent of the nitrogen occurs in the biosphere and the rest remains in the lithosphere (Stevenson, 1972). Nitrogen occurs in molecular form in the atmosphere. Merely, 0.04 per cent of the nitrogen occurs in compounds that are accessible to living beings. In the terrestrial system, nitrogen accounts for 4% of the biomass and 96% of soil organic matter (Rosswall, 1976). Despite the unavailability of 99.96% of the nitrogen gas, it is one of the indispensable elements that limit the primary production on which life forms depend for their daily requirements. It is a component of the proteins required for plant, animal and human life (Rosswall, 1976). An important feature of our planet ecosystem is the nitrogen cycle, since nitrogen availability affects the rate of major ecosystem processes, like primary production and decomposition. It is a mechanism by which nitrogen is converted into different chemical forms. Both biological and non-biological processes are associated with these transformations. Nitrogen cycle consists of nitrogen fixation, assimilation, mineralization, nitrification and denitrification (Rosswall, 1976). The process of conversion of molecular nitrogen to ammonia is known as nitrogen fixation. Some fixation occurs during lightning strikes; however most of the part is played by a group of free-living (eg. *Azotobacter*) or symbiotic

bacteria (*Rhizobium*, *Bradyrhizobium*, *Frankia* etc.). This fixed nitrogen is the primary source of available nitrogen in nature and is an essential step distributing the supply of this nutrient in the ecosystem (Saikia and Jain, 2007). Majority of plants and microorganisms incorporate ammonium salts as source of nitrogen either through glutamate dehydrogenase or glutamine synthetase/ glutamate synthase cycle (Miflin and Lea, 1977) by the process of assimilation. By the process of mineralization ammonium is released as molecular nitrogen. Some free living organisms like *Nitrobacter*, *Nitrosomonas* etc. utilize ammonium or nitrite compounds as an energy source (Ritchie and Nicholas, 1972). These are nitrifying organisms and the process is called nitrification (Rosswall, 1976). Denitrification reduces nitrates back to the largely inert nitrogen gas (N_2). Some bacterial species like *Pseudomonas* and *Clostridium* use nitrate as an electron acceptor in place of oxygen during respiration in anaerobic conditions and are denitrifiers. These denitrifiers complete the nitrogen cycle.

Abiological nitrogen fixation occurs by industrial synthesis, UV radiation and lightening. During industrial synthesis, Haber-Bosch process is used to produce nitrogenous fertilizers which are energy consuming and complicated (Ladha and Reddy, 1995). Nitrogen fertilizers are applied in agriculture in either nitrate (NO^{3-}) or ammonium (NH^{4+}) forms for enhancing crop production in order to meet the demands of the ever increasing world population. The chief constituent of the Green revolution of the 20th century was heavy dependency on chemical fertilizers, pesticides, and herbicides obtained from fossil fuels. Bumb and Baanante (1996) sketched an annual increase of about 1.2% in the use of nitrogenous fertilizer from 1990 to 2020. For the production of wheat, rice and maize alone 42 million tons of nitrogenous fertilizers are used annually on a global scale (Saikia and Jain, 2007). Approximately 25% of the nitrogen from nitrogenous fertilizers is lost by leakage, volatilization, denitrification and other factors resulting in high economic loss accompanied by severe environmental pollution. The high presence of nitrate and nitrite ions in water, food has resulted in diseases like cancer in new borns, respiratory problems, kidney failure etc. (Saikia and Jain, 2007). The ecosystem gets damaged owing to high levels of nitrate and ammonium ions in cultivated soils resulting in plant

toxicity (Saikia and Jain, 2007). In these circumstances, global concerns over the ill effects of chemical fertilizers, in addition to their rising costs for farmers in third world countries, it is necessary to reap the benefit of biological nitrogen fixation (BNF) based methodologies for cost effectiveness, maximizing output and obliterating the ill effects of environmental pollution.

Biological nitrogen fixation (BNF) is the mechanism of reduction of atmospheric dinitrogen to a metabolically active form exclusively by microorganisms. Biological nitrogen fixation is directly proportional to agricultural sustainability (Bohlool *et al.,* 1992). An understanding of the factors controlling BNF systems in the field is vital for the support and successful adoption in large scale. Gone are the concepts of trial-and-error experiments in agricultural sciences (Bohlool *et al.,* 1992). It is necessary to explore the possibilities for improving biological nitrogen fixation and its use by farmers on a global scale.

Biological nitrogen fixation is distributed among the domains of Bacteria and Archaea (Postgate, 1987; Young, 1992; Martinez-Romero, 2000). A broad range of microorganisms, spanning diverse taxonomic groups carry out the process of nitrogen fixation in a reaction catalyzed by nitrogenase enzymes (Kim *et al.,* 1992; Dean and Jacobson, 1992). Nitrogen fixing microorganisms are divided into free-living (fixing nitrogen independently) and symbiotic forms. Free living nitrogen fixing microorganisms include members of both Bacteria and Archaea. The symbiotic nitrogen fixers comprise of some members of cyanobacteria and genera like *Rhizobium*, *Bradyrhizobium*, *Sinorhizobium*, *Mesorhizobium*, *Azorhizobium* and *Frankia* (Paul and Clarke, 1996). Besides, some endophytic bacterial species like *Azoarcus* sp. too fix nitrogen in association with some tropical grasses. Free-living nitrogen fixing microorganisms comprise of *Azotobacter*, *Beijerinckia*, *Chlorobium*, *Clostridium*, some methanogenic archaeal members like *Methanococcus*, *Methanosarcina*, *Methanosprillium*, *Methanothermobacter* etc. Two types of symbiosis have been recognized amongst the symbiotic nitrogen fixers: one the legume-*Rhizobium* symbiosis and the other *Frankia*-actinorhizal symbiosis (Wall, 2000). The former involves more than 1700 plant species belonging to the families of Fabaceae,

Mimosae and Caesalpiniaceae while the latter involves 25 genera of dicotyledonous woody plants belonging to 8 families called actinorhizal plants (Bajwa, 2004). Symbiotic nitrogen fixing associations be it legume based or non-legume based result in the development of a specialized structure called root nodule where in the bacteria express the enzyme nitrogenase required for nitrogen fixation and fix nitrogen (Bajwa, 2004). Prin *et al.,* (1992) reported the presence of shoot nodules in some plants.

The complex process of rhizosphere colonization by bacteria, nitrogen fixation and export of fixed nitrogen is mediated by a number of enzymes and genes and requires large amount of energy (Saikia and Jain, 2007). Howard and Rees (1996) reported that thermodynamics of the biological nitrogen fixating mechanism limits its energetic requirements. During the process, the nitrogenase complex synthesized by the microorganisms is necessary. The ATP dependent reduction of nitrogen to ammonium is catalyzed by the nitrogenase complex and comprises of dinitrogenase and dinitrogenase reductase. Dinitrogenase exists structurally as a $\alpha_2ß_2$ tetramer having a molecular weight of 240 KDa and dinitrogenase reductase exists as a 60KDa homodimer (Howard and Rees, 1996). The total assembly of the nitrogenase complex requires many gene products. Dinitrogenase is encoded by *nif*D and K while *nif*H codes for dinitrogenase reductase (Dean *et al.,* 1993). Besides its function as a catalyst, dinitrogenase reductase function in the iron-molybdenum co-factor biosynthesis and dinitrogenase maturation (Dean *et al.,* 1993). The metallocluster known as iron molybdenum co-factor is the site for the binding of substrate and reduction of atmospheric nitrogen (Kim and Rees, 1992). Three major gene families' viz. *nif, nod* and *fix* are linked with biological nitrogen fixation. A number of *nif* genes have been identified in different microorganisms; some of them are *nif*E, *nif*N, *nif*W, *nif*V, *nif*B, *nif*S, *nif*A etc. (Fischer, 1994). These genes regulate the synthesis of the iron molybdenum co-factor as established by biochemical and genetic studies of *nif* mutants in *Klebsiella* (Filler *et al.,* 1986). The *fix* genes are the ones which are not frequent in free-living nitrogen fixers but play a major role in rhizobia legume symbiosis. (Fuhrman *et al.,* 1985; Gubler and Hennecke, 1986). These *fix* genes are located in a single operon and function

in transportation of electron to the nitrogenase system. The rhizobia-legume system houses the nodulation genes or *nod* genes. These genes are encoded in megaplasmids called symbiosis plasmids (Denarie and Debelle, 1996). Both *nod* genes and *nif* genes are closely located. Downie (1994) reported that rhizobial *nod* genes regulate the synthesis of *nod* factors. *Nod* genes respond to plant signals usually contain chemicals like flavones, flavanones and isoflavones (Peters *et al.*, 1986) exuding from the roots to the rhizosphere. The expression of *nod*D gene is influenced by the aforesaid chemicals and this expression facilitates the expression of other *nod* genes (Denarie and Debelle, 1996). In *Rhizobium*, *nod* ABCFELMN are involved in *nod* factor synthesis, while *nod* IJT encode proteins for *nod* factor transportation (Denarie and Debelle, 1996). Besides these genes, hydrogen uptake genes (*hup*) genes have been reported (Johansson *et al.*, 1983) in some microorganisms, which have the capacity to recycle hydrogen back to the nitrogenase complex thereby helping the plants to harvest the energy lost and improving the efficiency of nitrogen fixation (Johansson *et al.*, 1983).

The beginning of the Human genome project ushered a change in biology research (Ideker *et al.*, 2001) and pressed the accelerator for development of the science of bioinformatics. Bioinformatics is an interdisciplinary science comprising of theoretical and practical tools for realizing, creating and transmitting biological information. Bioinformatics in the broadest sense includes research in the genome structure, computational biology, genome expression, proteome analysis, genome and proteome engineering (Perez-Iratxeta *et al.*, 2006). With the development of PCR and establishment of the latest techniques of DNA sequencing there has been a scientific revolution in the understanding of the organism's genome. Arrival of genomics and proteomics has enabled the scientists to consider the cells as a system (Palsson, 2000). The complexities of the biological systems are now being studied with mathematical models and computer simulations (Palsson, 2000). Biology has been transformed into an informational science (Ideker *et al.*, 2001). The impact of this transformation is felt in the areas of medicine, agriculture, biological energy production and has made biotechnology a strong guiding force (Ideker *et al.*,

2001). The science of bioinformatics is associated with the challenge of decoding and studying huge number of genomic sequences present in databases. The knowledge obtained is used as a basis for understanding cell metabolism, physiology and allow one to design metabolic pathways and answer queries related to biology (Perez-Iratxeta *et al.*, 2006). The symbiosis of biotechnology and informatics is paying off. Bioinformatics tools offer the possibility of characterizing the house keeping genes which are essential for cellular functions and accessory gene pools associated with special biological functions that generate interest in the areas of medicine, agriculture and biotechnology (Perez-Iratxeta *et al.*, 2006).

Bioinformatics research has influenced the field of biological nitrogen fixation. The last several decades have witnessed a number of studies concerning the microbiology, genetics, biochemistry, physiology and ecology of the nitrogen fixing microorganisms. Since the last decade of the 20th century our understanding of the mechanism of nitrogen fixation has been given a new dimension with the advent of various techniques (Raymond *et al.*, 2004). Bioinformatics research in plant microbiology, constituting the associated plants and microorganisms is directed at acquiring the total nucleotide sequence of nitrogen fixing microorganisms and applying knowledge to successive post-genomic studies. The knowledge base upon which the genomics research, patents and social impact depend thrives on the complete genome sequence of nitrogen fixing microbes. Post genomic studies is likely to offer a global view of the expression, regulation, dynamics and evolution of the genomes from nitrogen fixing microbes and has the capability in offering new opportunities to preserve and improve biotic resources. In the early to mid 80's sequences of individual genes for *nif*, *nod* and *fix* for rhizobia started appearing in public domains but full scale genome sequencing started in 1997 when funds became available (Perret and Broughton, 2001). The availability of the sequences of nitrogen fixation related genes and proteins as also the whole genomes for a number of nitrogen fixing diazotrophs spanning different microbial groups has opened the possibility of analyzing their genomic context (Lakhsmi, 2007). Large scale sequencing has also offered a platform for crystallographers,

spectroscopists and more specifically computational biologists (Burley and Bonnano, 2002) aiming development of broad three-dimensional view for the structures of important proteins. This would present a novel outlook into the distribution of nitrogenase genes and their role in host metabolism as well as phylogeny. Genome projects of nitrogen fixing bacteria are no more a "rich man's sport'. The social benefit of the genome projects for nitrogen fixing microbes are great since they are the source of natural nitrogen in plants and soil. These projects have resulted in the availability of tremendous amount of biological data. This data includes information about genomes which in turn gives the idea about the proteome complement, proteins, codon usage etc.

Comparative genomics have emerged as one of the interesting areas of study. For comparing genomes of nitrogen fixing bacteria it is necessary to study their codon usage, analyze their proteomes, molecular phylogeny and protein. Studies of codon usage can be performed using parameters like GC content, GC3 content, relative synonymous codon usage (RSCU) (Peden, 1999), optimal codon anticodon energy (P2) (Peden, 1999), effective number of codons Nc (Wright, 1990), scaled chi-square (Peden, 1999), length of the amino acids (Laa) giving the number of translatable codons (Lloyd and Sharp, 1992), frequency of synonymous codons (Lsym) (Peden, 1999), codon adaptation index (CAI) (Sharp and Li, 1987), frequency of optimal codons (Fop) (Ikemura, 1981), codon bias index (CBI) (Bennetzen and Hall, 1982), hydrophobicity, aromaticity (Peden, 1999) 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 and Mrazek, 2000; Karlin *et al.*, 2001). There is dearth of comprehensive work on codon usage patterns in nitrogen fixing organisms. 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 *et al.,* 2003). In these perspectives, studies of codon usage patterns provide a platform for better understanding of the nature of nitrogen fixation taking into account the diverse nature of nitrogen fixing prokaryotes and interplay of the factors influencing codon usage. Besides, it is also necessary to study the nature of TTA containing codons in high GC rich nitrogen fixing microorganisms. TTA codons are one of the rarest codons found in bacteria (Li *et al.,* 2007). TTA codons apart from their normal leucine-coding function have been used as a regulatory device (Zaburannyy *et al.,* 2009). It is thus important to get an overview of the TTA containing genes in GC rich diazotrophs to find out steadiness in development of metabolic pathways that are TTA regulated.

Proper knowledge of the structure and function of proteomes are essential for understanding the underlying information (Fleming *et al.,* 2006). For proper functioning a protein should be stable and soluble (Das *et al.,* 2006; Jaenicke, 2000). Since, they are highly variable and complex their study in large numbers poses a significant challenge (Liu *et al.,* 2008). Most of the *in silico* studies concerning bacterial proteomes have stressed upon functional annotation and amino-acid composition with little focus on common physical properties (Knight *et al.,* 2004). Physical properties of proteins like isoelectric point (Nandi *et al.,* 2005) protein energetic cost (Akashi and Gojobori, 2002), and amino-acid adaptation index (Xia and Xie, 2001) are significant in portraying the functions of microbes. Although, some work has been done (Schwartz *et al.,* 2001; Knight *et al.,* 2004; Nandi *et al.,* 2005) in some microorganisms, very little has been done with nitrogen fixing proteomes. Consequently a broad picture on protein signatures for nitrogen fixing microorganisms is unavailable. Large scale sequencing has also offered a platform for crystallographers, spectroscopists and more specifically computational biologists using homology modeling techniques (Burley and Bonnano, 2002) aiming development of broad three-dimensional view for the structures of important proteins. The 3D structure of a protein is important for understanding protein function. In the homology modeling technique target sequences are at first retrieved. Then the template is selected and alignment of the target sequence with the template is

performed. Then the model is constructed and evaluation of the refined model is done using different techniques. The intrinsic dynamics of the protein models are studied using appropriate methods for getting insight into the functional properties of the protein (Centeno *et al.*, 2005).

Increasing number of sequences has resulted in diverse approaches for phylogenetic studies. Studies concerning the evolution of nitrogen fixation related genes has been highly debatable among researchers (Normand and Bousquet, 1989; Fani *et al.*, 2000; Zehr *et al.*, 2003; Henson *et al.*, 2004; Kechris *et al.*, 2006). Even though most of them rely on sequence alignment based techniques, there is a need for alternative methodologies to get a much clearer picture. Phylogenetic techniques based on sequence alignment and structures are inadequate in studying evolution since, sequence similarity becomes unreliable at identity levels below 25% (Mondal *et al.*, 2008). Some workers (Qi *et al.*, 2004; Sims *et al.*, 2008) have also pointed out fallacies in sequence alignment based methods. Alignment based methods fail to infer relationships from whole genome data (Qi *et al.*, 2004). Again, structure based techniques have incomplete span given that number of structures are inadequate to infer a conclusion. These studies leads one to assume that sequence based approach and structure based techniques are not fully adequate to throw light on the complex evolutionary process of nitrogen fixation. An alignment-free, condensed matrix method relying on nucleotide triplet based phylogeny (Randic *et al.*, 2001; Randic, 2000, Mondal *et al.*, 2008) has been used to re -examine the evolution of nitrogen fixation. The condensed matrix method of studying molecular phylogeny takes into account a set of invariants in a DNA sequence and determines the extent of resemblance among DNA sequences using the invariants (Randic *et al.*, 2001).

This PhD thesis has been developed at the Bioinformatics Facility, Department of Botany in the University of North Bengal. One of the main research interests were studies concerning nitrogen fixing microorganisms from the bioinformatics point of view.

Objectives of this work:

- Comprehensive comparative study of codon usage patterns of major nitrogen fixing microorganisms coming from different lineages with special reference to genes associated with nitrogen fixation. This would

help in understanding the major forces influencing the codon usage patterns in these organisms.

- Analysis of the nitrogen fixation related genes with respect to codon usage to understand their nature and interplay of different factors.
- Correlation of the codon usage bias with the tRNA content of the 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 predict the expression level of the genes.
- Correspondence analyses of codon usage and amino acid usage to investigate the major trends in codon and amino acid variations among the genes.
- Determination of the potentially highly expressed genes using CAI values, correlate the expression level of genes present in COG groups and study their influence on the lifestyle patterns of nitrogen fixing microbes.
- Determination of the TTA containing codons in high GC rich organisms, study their nature using different parameters and recognize their role in influencing the lifestyle of those organisms.
- Comparative analysis of proteomes for nitrogen fixing microorganisms using protein isoelectric point, amino acid adaptation index and protein energetic cost. This will help in recognizing characteristic features of protein functionality linked to nitrogen fixation and answer questions concerning the nature of their proteomes.
- Analysis of protein isoelectric point across COG groups to throw light upon the variation and lifestyle of these microorganisms.
- Development of a novel nucleotide triplet based condensed matrix method for analyzing molecular phylogeny of nitrogen fixing microorganisms. Characterization of the nitrogen fixation related 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 NifH proteins from *Frankia* using homology modeling

technique so as to study the structure-function relationship. Since, *Frankia* is a diazotroph that can exist as obligate symbiont as well as in free-living state a study of its NifH proteins can provide significant insights into their nature. NifH was considered owing to its importance in assisting electron transportation, FeMo cofactor biosynthesis and maturation of molybdo-ferro protein.