In-silico characterization of some nitrogenase proteins from symbiotic diazotrophs and cyanobacteria

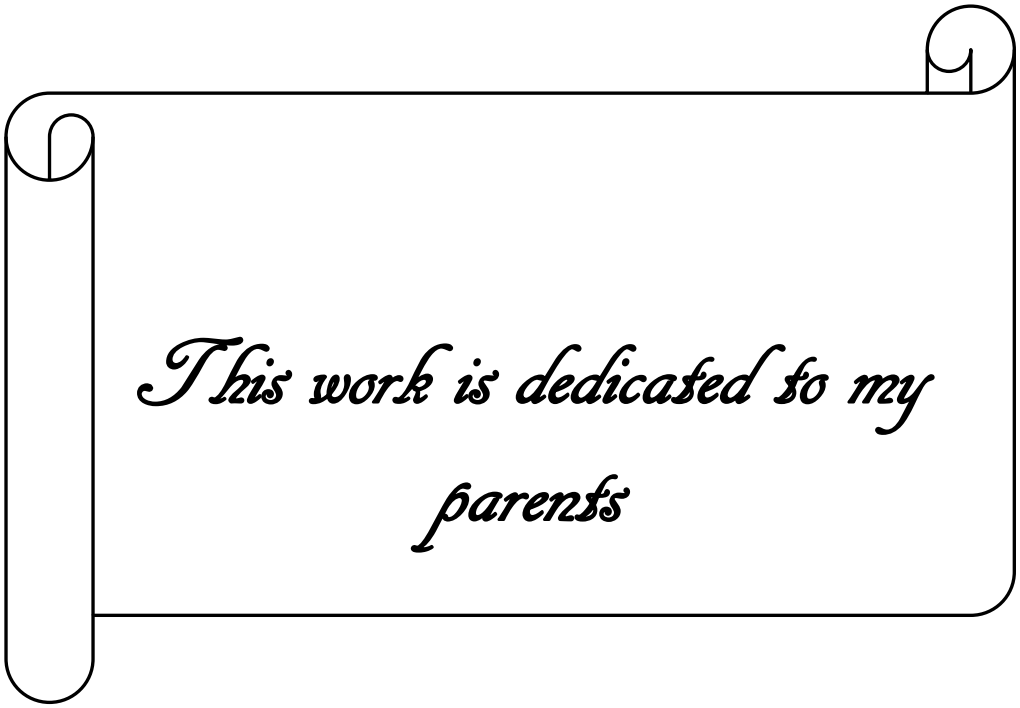
Thesis submitted to the University of North Bengal
For the Award of
Doctor of Philosophy
In
Botany

By
Subarna Thakur

Supervisor
Dr. Arnab Sen

Co-supervisor
Dr. Asim K. Bothra

Department of Botany
University of North Bengal
Raja Rammohunpur, Siliguri
March, 2013


This work is dedicated to my
parents

Declaration

I declare that the thesis entitled "*In-silico* characterization of some nitrogenase proteins from symbiotic diazotrophs and cyanobacteria" has been prepared by me under the supervision of Dr. Arnab Sen, Associate Professor of Department of Botany, University of North Bengal as Guide and Dr. Asim K. Bothra, Associate Professor of Department of Chemistry, Raiganj College (University College) as co-Guide. No part of this thesis has formed the basis for the award of any degree or fellowship previously.

[Subarna Thakur]
Department of Botany,
University of North Bengal
RajaRammohunpur, Siliguri

Date:

Certificate

We certify that Ms.Subarna Thakur has prepared the thesis entitled "*In-silico* characterization of some nitrogenase proteins from symbiotic diazotrophs and cyanobacteria", for the award of PhD degree of the University of North Bengal, under our guidance. She has carried out the work at the Department of Botany, University of North Bengal.

[Dr. Arnab Sen]
Supervisor
Department of Botany
University of North Bengal

Date:

[Dr. Asim K Bothra]
Co-Supervisor
Department of Chemistry
Raiganj College (University College)

Date:

Abstract

Nitrogen is an essential nutrient used by living organisms for survival. Molecular nitrogen or dinitrogen (N_2) makes up four-fifths of the atmosphere but is metabolically unavailable directly to higher plants or animals. It is available to some species of microorganism through Biological Nitrogen Fixation (BNF) in which atmospheric nitrogen is converted to ammonia by the enzyme nitrogenase. N -fixing microbes exhibit a wide diversity in terms of their habitat. Some are free-living whereas other form symbiotic associations with higher plants. Microorganisms that fix nitrogen are called diazotrophs.

Recent advances in next generation sequencing technology have ensured the availability of thousands of complete genomes of diazotrophs in public domain. But the arena of symbiotic diazotrophs and cyanobacteria is still relatively uncharted. Genomic approaches together with bioinformatics tools has opened a door for us to explore the complex molecular relationships that underpin symbiotic nitrogen fixation.

The present study deals with the characterization of proteins associated with nitrogen fixation in various symbiotic diazotroph and cyanobacteria using *in-silico* tools. Various aspects of these proteins were investigated on three fundamental levels i.e codon level, amino acid level and various structural level. For the codon-level study, whole genome sequences along with nitrogen fixing genes (NFGs) of symbiotic diazotroph and cyanobacteria were retrieved. The investigation into the codon usage pattern using parameters like GC, GC3, Nc, Fop etc revealed that diazotroph genomes and their nitrogen fixing genes are heterogeneous in nature. Different factors like mutational pressure, translational efficiency, GC3 compositional bias etc. play a vital role in shaping the codon usage variation. Rose plot analysis of codon usage revealed that AT-rich diazotroph genomes shows a preference for codons ending with A or T in third position whereas GC-rich genome show a preference for codons ending in G or C base. The evolutionary rates in

terms of rate of synonymous substitution (i.e. Ka/Ks) in whole genome along with NFGs in various nitrogen-fixing microbes were also looked upon. The evolutionary rate scrutiny revealed that most of the genes are under purifying selection but the rate of evolution is quite high in *Frankia* genomes in comparison to others. This result further lends support to the fact that *Frankia* genomes are experiencing turmoil in response to their adaptation to bio geographical surroundings. In sliding window analysis of Ka/Ks for *nif* genes, *nif*H was found to evolve slowly than others, indicating stronger stabilizing selection. The comparative genome analyses of Rhizobiales members and *Frankia* members have also been done to estimate the core and pan-genome. Pairwise BLASTP genome comparison was performed within and between genera to estimate the related gene homology. Genome comparisons have allowed the identification of a variety of similarities and differences in several groups of genes, including those involved in nodulation and nitrogen fixation.

Investigation in the amino acid level was initiated with the correspondence analysis on amino acid composition of proteomes. It showed that diazotrophs show striking signatures, with sharp segregations in proteome composition based on their lifestyle preferences. The rose plot and heat map of amino acid usage revealed considerable variation amongst nitrogen-fixing microbes. The next phase of the studies involved modeling of the nitrogenase proteins based on the homology modeling techniques. Three dimensional protein model of NifH protein was constructed for the symbiotic diazotroph, *Bradyrhizobium* ORS278 and cyanobacteria, *Anabaena variabilis*. The 3D models offered reliable insights into the 3D structural framework as well as structure-function relation of NifH protein. Metal binding sites and functionally important sites have also been identified. The model was further subjected to molecular dynamics simulation in aqueous environment. MD simulation of dimer provided great deal of information regarding the motional properties of these proteins. Nature of the flexible residues in context of their structural importance have been the focal point of this work.

Next level was to investigate the phylogenetic relationship of these proteins. Apart from the classical approach of sequence based phylogeny,

two different approaches have been utilized in this case. Results obtained for nucleotide triplet based phylogeny support evidence for polyphyletic origin, horizontal gene transfer and gene duplication for *nif*H, D and K phylogenies. The structure based phylogeny divulged about the evolutionary relationship of these protein with other related proteins and provided support to the fact that NifH protein and photosynthesis related light independent protochlorophyllidae reducatse (BchL) shared a common ancestor protein. Functional divergence analysis helped in identification of a number of vital amino acid sites which contribute to functional diversity. The investigation provides ample support to the fact that NifH protein and BchlL share robust structural similarities and have probably deviated from a common ancestor followed by divergence in functional properties probably due to gene duplication. Finally the structural analysis involving statistical coupling calculation for NifH/Bchl family helped in locating one sector comprising of 72 amino acids. The sector seems to be a global property of members of this family regardless of the subfamily architecture. Sectors represented features of protein structures than reflect the evolutionary histories of their conserved biological properties.

Preface

This Ph.D. thesis contains the result of research undertaken at the Department of Botany of University of North Bengal and Department of Chemistry, Raiganj College (University College) over the period from December 2008 to December 2012. Certainly, I would have never reached the point of finishing my dissertation without the help and support of others. These four years have been a challenging trip, with both ups and downs. Fortunately, I was not alone on this road, but accompanied by an extended team of experts, always willing to guide, inspire, help, and motivate me. For this, I would like to thank them.

Foremost, I take this opportunity to express my deep sense of gratitude and sincere thanks to my research supervisor, Dr. Arnab Sen, Associate Professor, Department of Botany, University of North Bengal for his invaluable guidance, support and persistent encouragement. No words can express my sincere and deep sense of reverence for him.

I am also indebted to my co-supervisor, Dr. Asim Kr. Bothra, Department of Chemistry, Raiganj College (University College) for his excellent guidance and critical suggestions all through my research work.

I am thankful to the Head of the Department and all the teachers and staff of the Department of Botany for their unequivocal support throughout.

I express my gratitude for all the members of Department of Chemistry, Raiganj College (University College) for supporting me to complete the work.

The work described in this thesis would not have been possible without the generous financial support from University Grants Commission and Council of Scientific & Industrial Research (CSIR), Government of India

I also appreciate Prof. Louis S Tisa, University of New Hampshire, USA and Prof. Phillipe Normand, University of Lyon1, France for their valuable suggestions. I express my gratitude to Prof. Rama

Ranganathan, University of Texas, USA and Prof. Ray Dixon, John Innes Centre, UK and Dr. Anirban Misra, Department of Chemistry, NBU, for providing me relevant information and materials to complete this work.

I am grateful to all my friends and Lab mates Arvind Kr. Goyal, Tanmayee Mishra, Saubashya Sur, Malay Bhattacharya, Debadin Bose, Uttam Mondol, BC Basistha, Ritu Rai, Manas Ranjan Saha, Sanghati Bhattacharya and Krishanu Ghosh for their support and encouragement all through my PhD tenure. On a personal note I would like to thank Ayan Roy for his support and constant backup particularly during those last few months of lab work and writing up.

It would be injustice if I neglect the services of my lab boys Poppai, Kurshid, Raj and Basu.

My sincere thanks to my parents, Mr Swapan Ranjan Thakur and Ms Sheuli Thakur, for their moral encouragement and constant motivation. My father has been my role model and silently supporting my perseverance for higher education. My mother has been my support system and she believed in me when others didn't. Lastly I am thankful to the almighty for all the blessings.

[Subarna Thakur]

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1
Introduction

The inception into the chronicle of nitrogen fixation has its beginning in 1772 with the discovery of the element nitrogen by Daniel Rutherford. It was found to be so inert that Antoine Lavoisier named it “azote”, meaning “without life”. Later it was established that nitrogen is the quintessential component of organic macromolecules like nucleic acid and protein, the building blocks of life form. All organisms require nitrogen to sustain and grow. The main reservoir of nitrogen is atmospheric N_2, which comprises approximately 78% of the atmosphere. Despite its abundance in the air we breathe, most of this nitrogen is unavailable for use by organisms. This is because the strong covalent triple bond between the N atoms in atomspheric N_2 molecules makes it relatively inert (Vitousek *et al.*, 2002). In fact, in order for plants and animals to be able to use nitrogen, N_2 gas must first be converted to more a chemically available form such as ammonium (NH_4+), nitrate (NO_3-), or organic nitrogen (e.g. urea $CO(NH_2)_2$). The process representing the transformation of nitrogen in nature is called the nitrogen cycle. Nitrogen cycle processes include assimilation, ammonification, nitrification, denitrification, anaerobic ammonium oxidation, and nitrogen fixation (Rosswall, 1976). One of the most important parts of the nitrogen cycle is nitrogen fixation. Nitrogen fixation is the process of conversion of atmospheric dintrogen (N_2) to the non-gaseous N compound ammonium (NH_4^+), which is usable by living organism. In nature, there are two major ways of fixing nitrogen. Natural abiotic N-fixation, which can be mediated by lightning or fires, oxidizes N_2 to nitrate (NO_3^-). The NO_3^- produced in this way can be washed out from the atmosphere with precipitation and is thus deposited in terrestrial ecosystems. The other way of nitrogen fixation involves activity of

certain soil bacteria that absorb atmospheric N_2 gas and convert it into ammonium. This process is known as Biological nitrogen fixation (BNF). Plants are able to absorb ammonia, nitrites, and nitrates from the soil and produce protein along with other N-containing biomolecules in the assimilation process. Animals do not have the enzyme required for catalyzing nitrogen fixation and consume plants as a source of nitrogen. Animals and plants can be decomposed in an ammonification process by decomposers, which results in the release of ammonia back into the environment. Soil ammonia converts to nitrites in a process called nitrification by nitrifying bacteria (Rosswall, 1976). Nitrites, which are useable for plants, can be converted to nitrates by nitrifying bacteria. Finally, nitrites return to the atmosphere through conversion to N2 by denitrifying bacteria in the denitrification process. 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. In anaerobic ammonium oxidation, nitrite and ammonium are converted directly into atmospheric N_2. This constant cycling means that the pool of fixed nitrogen within the biosphere is always replenished.

The inert nature of N_2 means that biologically available nitrogen is often in short supply in natural ecosystems, limiting plant growth and biomass accumulation. In fact, there is continuous loss of nitrogen from the reserve of combined or fixed nitrogen present in soil. It is continually depleted by such processes as microbial denitrification, soil erosion, leaching, chemical volatilization, and perhaps most important, removal of nitrogen-containing crop residues from the land. The nitrogen reserve of agricultural soils must therefore be replenished periodically in order to maintain an adequate (non-growth limiting) level for crop production. This replacement of soil nitrogen is generally accomplished by the addition of chemically fixed nitrogen in the form of commercial inorganic fertilizers or by the activity of biological nitrogen fixation (BNF) systems. The most important method of production of nitrogenous fertilizer is by the Haber process. This process requires high pressures and very high temperatures

which are achieved by utilizing vast amounts of fossil fuels as an energy source. These materials are non replaceable and, ultimately, exhaustible. BNF obtains the required energy from the oxidation of carbohydrates which have been formed by the photosynthetic activity of green plants. The energy for photosynthesis comes from sunlight. The energy for BNF is therefore indirectly derived from a universally available and inexhaustible source. The direct source of energy for BNF is therefore potentially available wherever conditions permit the growth of photosynthetic organisms. BNF contributes around 175 $TgNyear^{-1}$ to the global N-input of terrestrial ecosystem, hence it the primary source of N in soils and an eminently important process for the maintenance of life on earth (Paul and Clark, 1996). Although no eukaryotes are capable of biological di-nitrogen fixation by itself, nature has endowed several primitive micro-organisms (members of the domain Archea and Bacteria) with the capability to fix di-nitrogen biologically. Nitrogen fixing bacteria often form symbiotic relationships with host plants. This symbiosis is well-known to occur in the legume family of plants (e.g. beans, peas, and clover). In this relationship, nitrogen fixing bacteria inhabit legume root nodules and receive carbohydrates and a favorable environment from their host plant in exchange for some of the nitrogen they fix. The nodule also provides an environment with a low O_2 content, which is vital because nitrogenase, the enzyme that catalyzes nitrogen fixation is oxygen sensitive. Some oxygen however must be provided so that the bacteria can respire and produce energy required both for survival and to drive N_2 fixation. A special O_2-transporting protein, called leghemoglobin supplies this carefully controlled amount of O_2.

Symbiotic nitrogen fixing bacteria are represented by a phylogenetically disparate class of alpha- and beta-proteobacteria—usually collectively termed rhizobia—that have achieved the function of fixing atmospheric nitrogen (N_2) in symbiosis with legumes. The majority of the symbiotic species are represented in the alpha-proteobacteria order Rhizobiales, which, amongst many others, contain the agriculturally important nitrogen fixing genera of *Rhizobium, Bradyrhizobium, Mesorhizobium, Sinorhizobium (=Ensifer) and*

Azorhizobium. Amongst the actinobacteria, the genus *Frankia*, enters into symbiotic relationships with diverse set of woody non-leguminous dicot plants and was found to be nitrogen-fixing actinobacterium (Benson and Silvester, 1993). Numerous genera of non-leguminous angiosperms, such as *Alnus*, *Casuarina*, *Coriaria*, *Myrica*, etc., form root nodules in response to infection by the actinomycete *Frankia*. These associations between non-legumes and *Frankia* are called actinorhizal association. Such associations may achieve fixation rates as high as 100 kg N/ha/yr and are vital for maintaining the fertility of forest soil. Some gymnosperms, such as *Cycas*, *Macrozamia*, and *Podocarpus* , are capable of forming similar nitrogen fixing root nodule associations. There are also nitrogen fixing bacteria that exist without plant hosts, known as free -living nitrogen fixers. Free-living nitrogen fixing microorganism comprises of *Azotobacter, Beijerinckia, Cholorobium, Clostridium* and some methanogenic archea like *Methanococcus, Methanosarcina, Methanosprillium* etc. Besides these, there are many organisms that show various degrees of plant association, from truly endophytic life forms to strains that are merely enriched in the rhizosphere or on root surfaces (Kennedy and Islam, 2001). The latter are usually labeled “associative” diazotrophs. Bacteria such as *Azotobacter paspali, Azosprillium brasilense, Acetobacter diazotrophious, Herbasprillium* spp and *Azoarcus* spp. have been found to be in close association with roots of tropical grasses satisfying their fixed nitrogen requirement. Cyanobacteria are another group of ancient group of bacteria which is capable of nitrogen fixation. The photosynthetic cyanobacteria often live as free-living organisms in pioneer habitats such as desert soils or as symbionts with lichens. They also form symbiotic associations with other organisms such as the water fern *Azolla*, and cycads. The association with *Azolla*, where cyanobacteria (*Anabaena azollae*) been shown to be important for nitrogen inputs in rice fields. In aquatic environments particularly in ocean, cyanobacteria are generally assumed to be the major N_2-fixing microorganisms. The known open ocean N2-fixing cyanobacteria now fall into three major groups: (i) the filamentous non-heterocyst-forming

(heterocysts are specialized cells for N_2 fixation) like *Trichodesmium*, (ii) the filamentous heterocyst-forming symbionts with unicellular eukaryotic algae (e.g., *Richelia, Calothrix* and relatives), and (iii) single-celled or unicellular cyanobacteria (e.g., *Crocosphaera* and relatives of *Cyanothece*) (Zehr, 2011). The cyanobacterial genus *Trichodesmium* is alone responsible for ~42% of the total global nitrogen fixation of 240 Tg N_2 y^{-1} (Berman-Frank *et al.*, 2003).

Nitrogen fixation research and its application to agriculture have been ongoing for many centuries – from even before it was recognized as nitrogen fixation. The Romans developed the crop rotation system over 2000 years ago for maintaining and improving soil fertility with nitrogen-fixing legumes as an integral component. Even though crop rotation and the use of legumes was practiced widely since then, it was not until 1800 years later that insight came as to how legumes produced their beneficial effect. However, the science behind such practice was not revealed until Boussingault experimented with leguminous crops fixing N_2 in 1838; Hellriegel and Wilfarth showed definitive evidence for N2 fixation by microbes in legumes in 1886. The symbiotic relationship between legumes and nitrogen fixing bacteria is critical for agriculture, as it may have profound impacts on lowering costs for farmers, on land sustainability, on soil quality, and on mitigation of greenhouse gas emissions. The major importance of the symbioses is usually attributed to the decrease in the use of costly nitrogen based fertilizers, but the rehabilitation of infertile, environmentally stressed soils should also be highlighted. With an increasing global demand for food production, combined with the need to reduce carbon emissions, the reliance on biological nitrogen fixation as an alternative to nitrogen fertilizers is forecast to increase .The prospect of utilizing BNF for agricultural purposes has long been the driving force behind N-fixation research. The annual N-demand of the world's human population has been estimated at 23 Tg N $year^{-1}$, a figure that will increase considerably with the growing human population. The contemporary emphasis on sustainable development and the known negative side effects of mineral N-fertilizer (such as NO3-leaching into groundwater) have stimulated research in the field of BNF

throughout the past decades. This research has focused mainly on legumes and other symbiotic systems, which represent a well studied and widely used method of biological N-fertilization. In symbiotic systems, the ammonia produced by BNF is transferred directly to the plant where it is bound into organic molecules for transport (amino acid amides like asparagine and glutamine or ureides e.g. allantoin). The symbiosis provides the diazotroph with a steady influx of carbons sources and regulated oxygen supply which provides the basis for highly efficient N fixation. On the other hand, free-living diazotrophs will mostly incorporate the N they fix into their own biomass and its availability to plant will be indirect through subsequent mineralization of the biomass. Therefore, the application of free-living diazotrophs to soil as a method to promote plant growth has proven to be less successful in comparison to symbiotic system (Kennedy and Islam, 2001).

The biochemical machinery required for this process of biological nitrogen fixation is provided by the nitrogenase enzyme system. A number of biochemically and sequence-wise similar N_2-fixing enzymes are distributed among the various N_2-fixing bacteria. The first nitrogenase extract was prepared from the anaerobe *Clostridium pasteurianum* in 1960 and later nitrogenase has been extracted from various other diazotrophs. With few exception, all known diazotrophs contain nitrogenase that is a complex of two distinct metalloproteins, neither active without other. The complex comprises of Fe protein, product of *nif*H gene and the MoFe heterotetramer, products of *nif*DK genes (Rubio and Ludden, 2008). The Fe protein is a 64 kDa homodimer bridged by an intersubunit [4Fe–4S] cluster that serves as the obligate electron donor to the MoFe protein (Georgiadis *et al.*, 1992). The MoFe protein is a $\alpha_2\beta_2$ (around 250 kDa molecular mass) heterotetramer that houses the P-cluster, an [8Fe–7S] cluster that shuttles electrons to the FeMo-cofactor, a [Mo–7Fe–9S-homocitrate]cluster that provides the substrate reduction site (Kim and Rees, 1992). Although nitrogen fixation is a property of a phylogenetically diverse set of bacteria and cyanobacteria, in general, the sequences, structures, and functional properties of the nitrogenase Fe protein and MoFe protein are highly conserved between different organisms

(Howard and Rees, 1996). The total assembly of the nitrogenase complex requires many gene products. The *nif* HDK genes along with regulatory genes and accessory genes coding for coding for enzymes involved in electron transfer and metal cluster synthesis comprises the *nif* regulon (Dean and Jacobson, 1992). Adjacent to the *nif*H, *nif*D, and *nif*K genes in most *nif* operons are the *nif*E and *nif*N genes, which play a role in the biosynthesis of the nitrogenase metal clusters (FeMo-co). The NifE and NifN proteins have significant similarity to NifD and NifK, respectively, and are believed to have originated from an ancient duplication of a NifDK operon (Fani *et al.*, 2000). These core operons are supplemented by a number of other genes, which code for proteins involved electron transport (e.g. *nif*F and *nif*J in *Klebsiella pneumoniae*) or regulation (e.g. *nif*A) or in FeMo-cofactor synthesis (e.g. *nif*B, *nif*Q, *nif*V in *K. pneumonia* (Dean and Jacobson, 1992) and *A. vinelandii* (Rubio *et al.*, 2002). The transcription of the rhizobial genes coding for the nitrogenase complex is dependent on the *nif*A gene encoding a regulatory protein, which is expressed under specific environmental conditions, including low O_2 tension (Batut and Boistard, 1994). In addition to the standard nitrogenase system (*nif*DK, *nif*H) two important alternative systems have been identified and characterized. The main difference to the standard system is that these enzymes do not contain molybdenum (Mo). Instead, one of these alternative systems contains vanadium (*vnf*DK, *vnf*H) while the other contains only iron and no unusual metals (*anf*DK, *anf*H) (Eady, 1996). All three systems are highly homologous but do contain significant sequence differences. The alternative systems are transcriptionally regulated and in most studied organisms, *vnf* and *anf* are only expressed under conditions of Mo limitation (Bishop and Premakumar, 1992). Although these alternative nitrogenases could be derived paralogs of the Mo-requiring enzyme, they may instead represent primitive nitrogenases that have been maintained in several diverse lineages of prokaryotes (Anbar and Knoll, 2002). Of lately novel nitrogenase system was found in *Streptomyces thermoautotrophicus* (Ribbe *et al.*, 1997). The nitrogenase reductase in this system is a manganese superoxide dismutase that shows no homology to

*nif*H. Interestingly this Mo nitrogenase is not O_2-sensitive and its energy requirements are significantly lower than the classical nitrogenases (Ribbe *et al*., 1997).

Apart from *nif* genes, The "fix" and "nod" type genes are associated with BNF and nodule formation in rhizobial species, and many do not have homologues in the asymbiotic model organism *K. pneumonia* or other known free-living diazotrophs. The "fix" gene products are located in a single operon and are involved in electron transport to the nitrogenase holoenzyme and in regulation (Gubler and Hennecke, 1986). The *fix* genes represent a very heterogeneous class with *fix* L, *fix*J and *fix*K genes encode regulatory proteins whereas *fix*ABCX genes might code for an electron transport chain to nitrogenase. *fix*GHIS gene products are predicted to be transmembrane proteins, but their function in rhizobial nitrogen fixation is not defined properly (Fischer 1994), on the other hand, *fix*NOQP genes encode the membrane-bound cytochrome oxidase that is required for respiration of the rhizobia in low-oxygen environments (Delgado *et al.*, 1998). Along with *fix* genes, the rhizobia-legume system houses the nodulation genes or *nod* genes. These genes are generally housed in symbiosis-related megaplasmid close of the *nif* genes. Induction of nodulation genes leads to the production and secretion of return signals, the Nod Factors, which are lipochitooligosaccharides of variable structure (Denarie *et al.,* 1996). Nod factors are essential for the Rhizobiales to trigger root hair curling, to induce the formation of nodule primordia, and to enter the root via infection threads. NodD is the core signalling protein, reacting to plant flavonoids then binding to nod boxes, triggering the expression of a nod gene cascade. In *Rhizobium*, *nod* ABCFELMN are involved in *nod* factor synthesis, while *nod* IJT encode proteins for *nod* factor transportation (Denarie *et al*., 1996). Genome analysis of actinobacterial symbionts like *Frankia* have failed to permit identification of canonical *nod* genes but revealed other symbiosis-associated genes such as hydrogen uptake (*hup*) genes, sulfur-iron cluster (*suf*) gene and squalene-hopene cyclase (*shc*) gene which are found to be spread around the genomes (Kucho *et al.*, 2010). The hydrogen uptake genes has been reported to be involved in recycling hydrogen back to

nitrogenase complex thereby helping the plants to harvest the energy lost and improving the efficiency of nitrogen fixation (Brito *et al.,* 1997).

In the era of genomics, biological data are being produced at a phenomenal rate. The first complete genomic sequence obtained was that of *Haemophilus* influenza (Fleischmann *et al.,* 1995). Since this major achievement, many genomes have been sequenced and analyzed. These include members of three life domains: Bacteria, Archea and Eukarya. The major driving force for the development of genomics has been the completion of Human genome project in 2001. As a result of this surge in data, computers have become indispensable to biological research. Such an approach is ideal because of the ease with which computers can handle large quantities of data and probe the complex dynamics observed in nature. Thus, a new discipline, Bioinformatics, has emerged. Bioinformatics is defined as the application of computational techniques to understand and organize the information associated with biological macromolecules. Bioinformatics in the broadest sense includes research in the genome composition, genome expression, proteome analysis, genome and proteome engineering (Perez-Iratxeta *et al.,* 2007). Genome sequencing has provided a large amount of information on the genetic basis of nitrogen metabolism and its control in different diazotroph. The development of bioinformatics has opened a new window for providing novel understanding into the mechanism of diazotrophy. Thus, with the aid of bioinformatics, research related to nitrogen fixation has fully entered the era of genomics. Comparative genomics, together with functional studies, has led to a significant advance in this field over the past years The attainment of the nucleotide sequence of the cluster of *nif* genes in *Klebsiella pneumonia* can be considered as the precursor of genomic projects in the field of nitrogen-fixation (Arnold *et al.*, 1988). Formally, the first genome project with a N_2-fixing microorganism was the complete sequence of the symbiotic plasmid of *Rhizobium* sp. NGR234 (Freiberg *et al.*, 1996) , this was followed by sequencing of the symbiotic genomic compartments many other Rhizobiales strains. Completion of whole genome sequence of *M. loti* (Kaneko *et al.*, 2000), *S.*

meliloti (Galibert *et al.*, 2001) and *Bradyrhizobium japonicum* USDA110 (Kaneko *et al.*, 2002) proved to an important landmarks in the advancement of nitrogen fixation research. The comparison of symbiotic genome compartments with complete rhizobial genomes of different organisms has revealed interesting characteristics in regard to the organization and evolution of the genetic information necessary for the establishment of an efficient nitrogen-fixing symbiosis. In the era of genomics, more comprehensive approaches have become possible that allowed not only the identification of more genes necessary for root nodule formation and function, but which have also enabled their interactions to be studied on the whole-genome level. Large scale sequencing has also offered a platform for crystallographers, spectroscopists along with bioinformaticians to aim for development of broad three-dimensional view for the structures of important proteins. This would provide a novel outlook into understanding the working mechanism of the proteins actively involved in nitrogen-fixation and symbiosis. Genomics coupled with proteomics and trasncriptomics are now important high throughput techniques for qualifying and analyzing both gene and protein expression, discovering new gene or protein products, and perhaps these techniques hold the key for revealing the unknown facet of biological nitrogen fixation.

Availability of whole genome sequences of various diazotroph has given rise to the requirement of study of the relationship of genome structure and function across different biological species or strains. Comparative genomics is an attempt to take advantage of the information provided by the signatures of selection to understand the function and evolutionary processes that act on genomes. Comparative genomics represents a promising tool to reveal bacterial specificities in diazotrophic machinery. For comparing genomes of nitrogen fixing bacteria it is necessary to study their codon usage pattern, selective pressure shaping their life pattern, evolutionary pattern. The codon usage study provides information of use of different codons in a genome. Codon usage and codon preferences vary significantly within and between organisms and often it is seen that all codons are not used evenly. To dissect the patterns and

causality of codon usage, many indices have been proposed to measure the degree and direction of codon bias (Sharp and Li, 1987). Studies of codon usage can be performed using parameters like GC content, GC3 content, relative synonymous codon usage (RSCU) (Peden, 1999),effective number of codons Nc (Wright, 1990), frequency of optimal codons (Fop) (Ikemura, 1981), Among these indices, the codon adaptation index (CAI) was proposed as a measure of codon usage within a gene relative to a reference set of genes (usually ribosomal protein genes) (Sharp and Li, 1987). This index has been shown to correlate better with mRNA expression levels (Ikemura, 1981) and has been used to predict highly expressed genes in various organisms (Sharp and Li, 1987; dos Reis *et al.,* 2003; Martin-Galiano *et al*., 2004; Wu *et al*., 2005). 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 (Ziegel, 1993). This method can be used for extracting information regarding variability of amino acid usage in the proteome of the organism. It has been postulated that major trends in codon usage patterns across genomes are determined by computational bias, mutational pressure or translational selection in high or low G+C containing organisms (Knight *et al.,* 2001). 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 factors influencing codon usage.

Other than codon usage, molecular evolution of genes is another aspect which needs to be investigated. A more reliable index of genetic drift over evolutionary time is the ratio of Ka (nonsynonymous substitutions per site) to Ks (synonymous substitutions per site) for a large set of genes, based on comparisons of related species. The terms Ka/Ks and dN/dS are often used interchangeably. But for the computational purpose, the ratio of nonsynonymous (Ka) to synonymous (Ks) nucleotide substitution rates is frequently used as an indicator of selective pressures on protein-coding genes. The Ka/Ks ratio reflects the rate of adaptive evolution against the background rate (Hurst, 2002). This

parameter has been widely studied in the analysis of adaptive molecular evolution, and is regarded as a general method of measuring the rate of sequence evolution. To study the impact of selective pressures on the pattern of genetic divergence between diazotroph, it is necessary to find out the pairwise ratio of Ka/Ks between the orthologous genes. Oftentimes, different regions of a single gene can be exposed to different selective pressures (Hurst and Pal, 2001). In these cases, calculating Ka/Ks over the entire length of the gene does not provide a detailed picture of the evolutionary constraints associated with the gene. Hence, a finer analysis of the Ka/Ks ratio by using sliding windows of different sizes provides minute details of the selective constraints of the nitrogen fixation related genes. Besides these, other vital aspect of comparative genomics is the pangenome and core genome construction in various nitrogen-fixing microbes. It is a data-driven approach towards elucidating the differences as well as similarities between closely related genomes. Pangenome concept has been introduced (Medini *et al.*, 2005) as the sum of "core" (conserved in all strains) and "accessory" (variable among strains) genes. This provides a theoretical representation of representations of a collective and a conserved protein pool, respectively. The proteomes can also be compared by the blast matrix method where pairwise BLASTP genome comparisons are performed within and between genera. Informative BLAST Atlases can be constructed to visualize genomic variation within genera. These approaches of genome and proteome assessment would be helpful in elucidating the size and the functionalities of the symbiotic diazotroph pangenome especially at the level of the symbiotic diversity.

To have a better understanding of the nitrogen fixing mechanism, we have to look into the structural aspects of the nitrogenase components. With recent developments in the field of proteomics, a large amount of biological data in the form of amino acid sequences of nitrogenase proteins from a wide range of microbes are now available in the databases. However, very little is known about the 3D structure and role of all these proteins. Cystallographic structure of nitrogenase molybdenum-iron protein of diazotrophs like *Azotobacter vinelandii* (Kim and Rees, 1992) and

Clostridium pasteurianum (Kim *et al.*, 1993) has been resolved. However, the obstacles to studying nitrogenase are multiple and varied: they include the extreme oxygen sensitivity of the proteins; the size and complexity of the iron-sulfur containing metallocenters and difficulties in obtaining crystals of nitrogen bound to nitrogenase (Rees and Howard, 2000). One successful way to learn something about the protein is to identify its homologs and use information from those homologs to annotate the protein of interest. So, modelling of proteins based on homology is a dependable procedure that can predict the 3D structure of a protein with correctness comparable to that determined at low resolution by experimental methods (Marti-Renom *et al.*, 2000). In the homology modelling technique target sequences are first retrieved. Then a suitable template is selected and the template is aligned with the target sequence. Based on this alignment, 3D models of the protein are constructed. The three-dimensional models constructed by the homology modelling technique often contain certain amount of errors and it becomes 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 refinement process, the protein models are 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. Finally the models are evaluated of their internal quality and reliability. The conformational and molecular dynamics of the protein are studied using the molecular simulation program for getting insight into the motional properties of the proteins. In molecular dynamics atoms and molecules are allowed to interact over time at a given temperature following the laws of classical mechanics and which provides a detailed description of atomic motion (Dodson *et al.,* 2008). Dynamics trajectory is analyzed to get several important information like Radius of gyration, Root mean square deviation (RSMD), Root mean square fluctuation (RMSF), and also Principle component analysis. The RMSD was used to evaluate the deviation of the structure from the original starting structure over the

period of the simulation whereas RMSF is a measure of flexibility of each residue over the trajectory. The Radius of gyration provides an idea about the compactness of the structure during the simulation.

In addition to molecular dynamics studies, another field which needs attention is the regulation of the nitrogenase system particularly due to post translational modification. As seen in many photosynthetic nitrogen-fixing bacteria like *Rhodosprillium* and *Azospirillum* this involves reversible ADP ribosylation of dinitrogenase reductase (NifH) at arginine residue which remains in close proximity of the Fe4S4 cluster in presence of the enzyme dinitrogenase reductase ADP ribosyltransferase (DRAT) (Lowery and Ludden, 1989). The ADP ribosylation prevents the productive association of dinitrogenase reductase with dinitrogenase, such that neither electron transfer nor ATP hydrolysis can occur (Ma and Ludden, 2001). In these circumstances, the nitrogenase activity is said to be switched off. The modification site i.e Gly-Arg(ADR-ribose)-Gly-Val-Ile-Thr is highly conserved in nitrogenase iron protein of other nitrogen-fixing organism (Pope *et al.,* 1985) but its role has not yet been probed in symbiotic diazotrophs. *In silico* docking procedures utilizing the three dimensional protein model of NifH protein can be implemented to investigate the role of arginine residue in binding of the ligand ADP-Ribose molecule and structural changes induced in the protein due to the ligand binding.

With the advancement of genome sequencing, there are now hundreds of published nitrogenase gene sequences from many different species, especially for *nif*H as this gene have been widely used for diversity studies. Sequence based phylogeny of NifH and in some cases NifD proteins have been extensively used to infer about the phylogenetic studies of nitrogen fixing organisms (Normand and Bousquet, 1989; Fani *et al.*, 2000; Zehr *et al.*, 2003). But the exact course and time of evolution of nitrogenase complex has been a matter of much debate and has not been fully elucidated. More over the nitrogenase iron protein (NifH) have been found to posses some distant relatives. Some of these are well-known proteins involved in the synthesis of photosynthetic pigments, namely protochlorophyllide reductase (BchL or ChlL) and chlorin reductase

(BchX) (Burke *et al.*, 1993). Therefore, to fully understand the emergence and evolution of nitrogen fixation ability (diazotrophy) among Prokaryotes, we have to take into consideration the ancient descendants of the protein. In such cases, phylogenetic techniques based on sequence alignment are inadequate in studying evolution since; sequence similarity becomes unreliable at identity levels below 25%. Alternative phylogenetic techniques like condensed matrix method relying on nucleotide triplet based phylogeny has been used to examine the evolutionary pattern of nitrogenase proteins (Sur *et al.,* 2010), but these methods are still rudimentary and are not fully adequate to throw light in complex evolutionary process of biological nitrogen fixation. For proteins sharing relatively low sequence identity as in this case, 3D structures are better than primary sequences for modeling of protein evolution (Balaji and Srinivasan, 2007). Thus, structure-based phylogenetic approach can also be applied to increase the reliability of evolutionary reconstructions of nitrogenase system. Functional divergence in NifH/ Bchl family leading to the shift in protein function from that of its ancestor is another important aspect which needs to be investigated. The gene duplication event within a gene family often leads to functional divergence and as a consequence of which some residues encounter altered functional constraints (Landgraf *et al.*, 1999). The evolutionary rates at these sites will vary in different homologous genes of a gene family and this is known as type I functional divergence (Landgraf *et al.*, 1999; Dermitzakis and Clark, 2001). Analysis of functional divergence in Nifh/Bchl protein family will elucidate intrinsic functional diversity, and underlying evolutionary mechanism among these proteins. Moreover, functional divergence among proteins is often assumed to be strongly influenced by natural selection particularly positive selection (Tennessen, 2008). Maximum likelihood (ML) tests are often implemented to investigate presence of positive selection at individual amino acid sites.

Apart from these, another crucial aspect of structural studies is related to the structural organization of protein. Recently, concept of a new protein structural organization unit called “Sectors” have been introduced which

is distinct from the traditional hierarchy of primary, secondary, tertiary, and quaternary structure (Halabi *et al.*, 2009). This is based on statistical analysis of correlated evolution between amino acids. Each sector is physically connected in the tertiary structure, has a distinct functional role, and constitutes an independent mode of sequence divergence in the protein family. Sectors are regarded as features of protein structures than reflect the evolutionary histories of their conserved biological properties (Halabi *et al.*, 2009). Sector characterization based on statistical coupling analysis has already been implemented in many protein families including chaperons (Russ and Ranganathan, 2010; Smock *et al.*, 2010). In NifH protein family, sector analysis can provide enhanced understanding of their biological properties and evolutionary origins.

This PhD thesis has been envisaged and developed at the Bioinformatics Facility, Department of Botany in the University of North Bengal. The main research focus of this work is the characterization of the nitrogen fixing machinery of diazotrophs particularly the proteins involved through the use of various bioinformatics tools.

Objectives of this work:

- Comprehensive comparative study of codon usage pattern variation in various symbiotic nitrogen fixing bacteria and cyanobacteria, with special reference to genes associated with nitrogen fixation. This would help in understanding the major forces influencing the codon usage pattern in these organisms. Correspondence analysis of amino acid usage will also be done to probe the apparent causes behind the functional adaptations of encoded proteins.
- Analysis of Codon adaptation index. It is a measurement of relative adapted-ness of a gene's codon usage towards the codon usage of highly expressed genes. The CAI values will be used to evaluate the expression level of the nitrogen-fixation related genes. The CAI values were further correlated to the experimental microarray expression data obtained for the nitrogen-fixing actinomycetes *Frankia*.
- Investigation of selective pressures shaping the genome evolution of various symbiotic diazotroph and cyanobacteria. The rates of non-synonymous to synonymous substitutions or Ka/Ks ratio were assessed in orthologus proteins particularly those involved in nitrogen-

fixation in different microbial strains. The Ka/Ks ratio is widely used as an indicator of the extent of purifying selection acting to conserve coding sequences. It provides an idea about the evolutionary rate of protein coding genes.

- Sliding window analysis of the site -specific evolutionary constraints amongst the homologous coding sequences of genes encoding the components of nitrogenase enzyme apparatus.
- Pan-genome and Core-genome construction of various Rhizobiales strains and *Frankia* strains. In this study, the genomes of selected strains were compared by different comparative genomics tools to find a genomic interpretation to the phenotypic differences and to define a set of accessory genetic factors related to the symbiotic process and subsequent nitrogen fixation, as well as to obtain a better understanding of the strategies adopted by disparate strains to maintain their symbiotic apparatus.
- Secondary structure prediction along with determination of three dimensional structure of nitrogenase Fe protein (NifH) of the symbiotic bacterium *Bradyrhizobium sp.* ORS278 and the cyanobacteria *Anabaena variabilis* utilizing the homology modelling technique. The nitrogenase iron protein from two varied nitrogen-fixing bacteria coming from different lineages was selected for assessment of their structural features. The nitrogenase iron protein was considered owing to its importance in nitrogenase system particularly in electron transportation, FeMo-cofactor biosynthesis and maturation of the molybdo-ferro protein. The functionally relevant regions and position of 4Fe-4S cluster in the protein were also investigated.
- Molecular dynamic simulation of suitable refined model of the NifH protein to understand the motional and structural properties. The trajectory files obtained during the whole simulation run was analyzed and was interpreted to understand the structure-function relationship in context of the biological activity of the protein.
- Utilization of the 3D model of the NifH protein from *Bradyrhizobium sp.* ORS278 in *in-silico* docking experiment to highlight the importance of covalent modification site in post-translational regulation of the nitrogenase enzyme.
- Application of a novel structure based phylogenetic method to assess

the evolutionary relationship of the nitrogenase protein with other related proteins and compare its results with those of amino acid sequence based and codon-triplet based phylogeny.

- Estimation of functional divergence in the NifH/BchL family by conducting pair-wise functional divergence analysis between *nif*H paralogous genes. A number of protein domains and amino acid residues which contribute to predicted functional divergence have been identified. The maximum likelihood tests were utilized for detection of positive selection at the amino acid level.
- Use of the statistical coupling analysis (SCA) approach for characterizing the pattern of evolutionary constraints on and between amino acid positions in NifH protein family. This technique will help in correlating the conservation of amino acid with the architecture of functional interactions between amino acids and more importantly, help in defining the physical principles underlying protein structure and function.

Chapter 2
Review of Literature

2.1 Biological nitrogen fixation: an overview

Nitrogen is an essential nutrient used by living organisms for survival. It is the nutrient that is most commonly deficient, contributing to reduced agricultural yields throughout the world. Molecular nitrogen or dinitrogen (N_2) makes up four-fifths of the atmosphere but is metabolically unavailable directly to higher plants or animals. It is available to some species of microorganism through Biological Nitrogen Fixation (BNF) in which atmospheric nitrogen is converted to ammonia by the enzyme nitrogenase. The ammonia is then transferred to the higher plant to meet its nutritional needs for the synthesis of proteins, enzymes, nucleic acids, chlorophyll, etc and subsequently it enters the food chain. Thus all eukaryotes (including higher plants and animals) naturally depend on the BNF activity of the N-fixing microbes for their N supply. Microorganisms that fix nitrogen are called diazotrophs. According to current knowledge, only prokaryotes (members of the domains Archaea and Bacteria) are capable of performing BNF. The ability to fix nitrogen is widely, though paraphyletically, distributed across both the bacterial and archaeal domains (Raymond *et al.,* 2004). There are two types of diazotrophic prokaryotes; those that are free-living and those that form symbiotic relationships. Those microbes, such as *Azotobacter* and *Clostridium* that fix nitrogen independent of other organisms, are called free living. The free-living diazotrophs require a chemical energy source if they are nonphotosynthetic, whereas the photosynthetic diazotrophs, such as the cyanobacteria, utilize light energy (Leigh, 2002). Some diazotrophs called rhizobia enters into symbiotic relation with legumes like clover and soyabean. The symbiosis between legumes and the nitrogen-fixing rhizobia occurs within

nodules mainly on the root and in a few cases on the stem (Burns and Hardy, 1975). A similar symbiosis occurs between a number of woody plant species and the diazotrophic actinomycete *Frankia* (Pedrosa *et al.*, 2000). These symbiotic associations are the greatest contributions of fixed nitrogen to agricultural systems.

Exploitation of biological nitrogen fixation for agricultural benefits has long been sought after. BNF provides a means to meet the needs of a growing population with a nutritious, environmentally friendly, sustainable food supply. This makes the need for BNF research very compelling in the current scenario. In the last two decades, many exciting discoveries in nitrogen fixation have been reported—genomes have been sequenced, the "omics" approaches have been applied to both symbionts, new genetically modified crops have become commonplace in agriculture. Biochemical research into the workings of nitrogen fixation is generally focused on the enzyme complex called nitrogenase. Other than its usual function, this system has emerged as a model for more general biochemical processes, such as signal transduction, protein-protein interaction, inter- and intramolecular electron transfer, complex metal cluster involvement in enzymatic catalysis etc (Peters *et al.,* 1995).

2.2 Various aspects of biological nitrogen fixation:

2.2.1 The ecological importance of nitrogen-fixation:

The circulation of nitrogen through biological and physical world is an imperative feature of our planetary ecosystem. The nitrogen cycle is of particular interest to ecologists because nitrogen availability can affect the rate of key ecosystem processes, including primary production and decomposition. Since N is an essential nutrient for all living organisms, N-fixation is a key process of the N-cycle. Nitrogen fixation particularly those carried out by prokaryotes is ecologically important as an input of fixed nitrogen (N) in many terrestrial and aquatic ecosystems. The biologically fixed nitrogen in the form of ammonia relieves ecosystem N limitation. Majority of plants and microbes incorporate ammonium salts as source of nitrogen either through glutamate dehydrogenase or glutamate synthetase cycle by the process of assimilation (Miflin and Lea, 1976). The animal in turn fulfils their nitrogen requirement

through the food chain. After metabolic cycling through the biosphere, fixed N is returned to the atmospheric reservoir through the process of denitrification. The balance between N_2 fixation and the reverse process, denitrification, can determine the net biologically available N for the biosphere (Arp, 2000). The ecological control of biological nitrogen fixation has been extensively studied by Vitousek *et al.*, (2002), in various ecological system. N2 fixation often found to be interacting with other biogeochemical cycles to control the N status of the ecosystem. Ecological studies by carried out Cleveland *et al.,* (1999) points to the fact that importance of BNF in a particular ecosystem depends on the systems nutrient status, which is in turn linked to the stage of ecosystem development. In early successional stages of soil development phototrophic diazotrophs may be important colonizers of oligotrophic soil surfaces. BNF by free -living heterotrophs might be especially important in systems with high accumulations of carbon sources, e.g. in forest soils with high inputs of litter and decaying plant debris (Giller and Day, 1985). On the other hand, in ecological systems with high concentrations of mineral N or other unfavorable conditions, BNF will contribute little to the N-budget (Cleveland *et al.,* 1999). Besides the terrestrial ecosystem, biological nitrogen fixation has gained recognition as an important source of nitrogen for supporting oceanic primary production. The role of BNF in marine nitrogen cycle has been extensively discussed in the works of several researchers (Fogg, 1978; Lipschultz and Owens, 1996; Falkowski 1997; Capone, 2001) and more recently by Rijkenberg (2011). Nevertheless, knowledge of the diversity, abundance and distribution of marine diazotrophs (N2 fixing micro -organisms) is limited and the factors that control N2 fixation in the marine environment are still poorly understood.

2.2.2 Biological N_2 fixation and sustainable agriculture:

Natural reserves of soil nitrogen are normally low, so commercially prepared N fertilizers must be added to increase plant growth and vigor. Chemical fertilizers had a substantial impact on food production in the recent past, and are today an indispensable part of modern agricultural practices. But for the farmers of developing countries N fertilizers are neither

affordable nor widely available. Moreover, the harmful effects on the environment of heavy use of N fertilizer are becoming more evident day by day. Further, the fossil fuels which are used in the production of N fertilizer are becoming scarcer and more expensive. At the same time, the demand for food is going up as populations increase. Therefore, there is a great need to search for all possible avenues. The process of biological nitrogen fixation offers economically attractive and ecologically sound means of reducing external nitrogen input and improving the quality and quantity of internal resources. Biological nitrogen fixation is the reduction of atmospheric N2 gas to biologically available ammonium, mediated by prokaryotic organisms in symbiotic relationships, associative relationships, and under free-living conditions (Postgate, 1998). The fixed nitrogen that is provided by biological nitrogen fixation is less prone to leaching and volatilization and therefore the biological process contributes an important and sustainable input into agriculture. Nitrogen input through BNF can help maintain soil N reserves as well as substitute for N fertilizer to attain large crop yields (Peoples and Craswell, 1992). An understanding of the factors controlling BNF systems in the field is vital for the support and successful adoption in large scale in an agricultural context.

Wani *et al*., (1995) highlighted the importance of biological nitrogen fixation of legumes in sustainable agriculture in semi arid tropic region. Legumes, one of the most important plant families in agriculture are often involved in a remarkable symbiosis with nitrogen fixing rhizobia. Legumes are often considered to be the major nitrogen-fixing systems, as they may derive up to 90% of their required nitrogen from N_2. The quantity of atmospheric N fixed through forage legume biological N fixation can range as high as 200 kg/ha per year (Peoples *et al*., 1995). The symbiotic association of actinorhizal species helps in improving soil fertility in disturbed sites such as eroded areas, sand dunes, moraines etc. Actinorhizal plant nitrogen fixation rates are comparable to those found in legumes (Torrey and Tjepkema, 1979; Dawson, 1983). Nitrogen fixing *Azolla*-Cyanobacteria symbiosis has been widely used used to enrich rice paddies with organic nitrogen in Southeast Asian countries

like China, Vietnam, and Southeast Asia (Watanabe and Liu, 1992). The rice paddies of Asia, which feed over half of the world's population, depend upon cyanobacterial N2 fixation (Irisarri *et al*., 2001).

2.2.3 N_2 fixation and silvicultural systems:

Biological nitrogen fixation occurs in most forest ecosystems and thus is a vital factor in all silvicultural systems. Nitrogen added by N-fixing plants could be important in replacing nitrogen losses from timber harvesting or fire. Symbiotic N fixation particularly those involving actinorhizal trees has the potential to add significant amounts of N to forest soils. Actinorhiza is the result of a symbiotic relationship between a special soil N2-fixing actinomycete (mostly the genus *Frankia)* and fine plant roots. Since the result of the symbiotic association is a root nodule, the host plant is known as an actinomycete-nodulated plant or actinorhizal plant (Torrey, 1978). Actinorhizal associations range from the arctic to the tropics and from the semi-desert to rainforest ecosystems. Many actinorhizal plants are pioneering species, like *Alnus*, which grows in moist environments, or *Myrica*, which grows on landslides, eroded slopes, and mined areas, or Casuarinas, which even grows in nutrient deficient sandy loams. At the forest-stand level, the rates of symbiotic N2 fixation reported for actinorhizal trees range between 12 and 320 kg N ha^{-1} yr^{-1} (Wheeler and Miller, 1990). By comparison, asymbiotic N fixation by free-living soil prokaryotes typically contributes only 1 kg N ha^{-1} yr^{-1} or less in forest ecosystems (Jurgensen *et al*., 1992) whereas associative N2 fixation has been suggested to fix up to 50 kg ha^{-1} yr^{-1} in the rhizosphere of conifer roots (Bormann *et al*., 1993). According to Myrold and Huss-Danell (2003), actinorhizal plant like Alders, are beneficial at improving nutrient-poor soils. The annual input of N from N_2 fixation in alder stands ranges from 20 to 300 kg N/ha, depending on stand age, stand density and site conditions. According to Dommergues (1997), the nitrogen-fixing potential of a number of actinorhizal plants, e.g., *Casuarina* sp. and *Alnus* sp., is high but the amount of N actually fixed in the field is often low because the expression of this potential is limited by either unfavorable environmental conditions or improper management practices.

2.2.3 Physiological and phylogenetic diversity of diazotrophs

Farmers have known, probably since the time of the Egyptians, that legume such as pea, lentil, and clover are important for soil fertility. The practice of crop rotation, intercropping and green manuring were extensively described by then Romans, but it was not until the 19th century that an explanation for the success of the legumes in restoring soil fecundity was uncovered. The discovery of nitrogen fixation was attributed to the German scientists Hellriegel and Wilfarth, who in 1886 reported that legumes bearing root nodules could use gaseous nitrogen. Shortly afterwards, in 1888, Beijerinck, a Dutch microbiologist, succeeded in isolating a bacterial strain from root nodules. This isolate happened to be a *Rhizobium leguminosarum strain* (Franche *et al*., 2009). Beijerinck (in 1901) and Lipman (in 1903) were responsible for isolation of *Azotobacter* spp., while Winodgradsky (in 1901) isolated the first strain of *Clostridium pasteurianum* (Stewart, 1969). Discovery of nitrogen fixation in blue-green algae was established much later (Stewart, 1969). The identification of nitrogen-fixing microbe from root nodules of non-leguminous plants like Alder generated considerable controversy for a while. It was Brunchorst, who named the microbe *Frankia subtilis* (Pawlowski, 2009). Hiltner (1898) recognized the nodule inhabitant as an actinomycete, Gram-positive bacteria closely related to *Streptomyces* (Hirsch, 2009). Pommer (1959) was probably the first person to obtain an isolate, but it did not reinfect its host plant. For a long time, diazotrophy in the actinomycetes was thought to be limited to the genus *Frankia* but through the years several other actinomycetes have been shown to have *nif* genes (Gtari *et al*., 2012). Over the years there have been continual discoveries of new diazotrophs, revealing that this function is performed by a very diverse group of prokaryotes. In the last decades, the use of molecular technologies for the direct detection of the genes of biological nitrogen fixation has shown that the capacity for diazotrophy is even more wide-spread than previously expected.

Although nitrogen fixation is not found in eukaryotes, it is widely distributed among the Bacteria and the Archaea, revealing considerable biodiversity among diazotrophic organisms. The

ability to fix nitrogen is found in most bacterial phylogenetic groups, including green sulphur bacteria, Firmibacteria, actinomycetes, cyanobacteria and all subdivisions of the Proteobacteria. In Archaea, nitrogen fixation is mainly restricted to methanogens. The ability to fix nitrogen is compatible with a wide range of physiologies including: aerobic (e.g., *Azotobacter*), facultatively anaerobic (e.g., *Klebsiella*) or anaerobic (e.g., *Clostridium*) heterotrophs; anoxygenic (e.g., *Rhodobacter*) or oxygenic (e.g., *Anabaena*) phototrophs; and chemolithotrophs (e.g., *Alcaligenes, Thiobacillus, Methanosarcina*) (Young, 1992). Diazotrophs show considerable diversity in terms of habitats. They are found as free-living in soils and water, associative symbioses with grasses, actinorhizal associations with woody plants, cyanobacterial symbioses with various plants. The most widely known and discussed feature of diazotrophs is their symbiotic association with a number of leguminous plants collectively referred to as Rhizobia. The rhizobia are Gram-negative and belong to the large and important Proteobacteria division and include the genera like *Agrobacterium, Allorhizobium, Azorhizobium, Bradyrhizobium, Mesorhizobium, Rhizobium, Sinorhizobium, Devosia, Methylobacterium, Ochrobactrum* etc. (Franche *et al*., 2009). These soil bacteria are able to invade legume roots in nitrogen-limiting environments, leading to the formation of a highly specialized organ, the root nodule. These specialized root structures offer an ecological niche for the microbe to fix nitrogen (Mylona *et al*., 1995). Nodule formation is a complex process that requires a continuous and adequate signal exchange between the plant and the bacteria. Rhizobia are attracted by root exudates and colonize the plant root surface. Specific flavonoid compounds present in the exudates activate the expression of the bacterial nodulation (*nod*) genes involved in the synthesis and secretion of Nod-factors (NF), lipochito-oligosaccharides that are recognized by the plant. Nod factors, together with additional microbial signals such as polysaccharides and secreted protein, allow bacteria attached to root hairs to penetrate the root (Schultze and Kondorosi, 1998). The tip of a root hair, to which rhizobia are bound, curls back on itself, trapping the bacteria within a pocket of host cell

wall (Brewin, 1993). This is followed by formation of infection threads by the rhizobia allowing them to move between cells. The formation of infection threads triggers cell division in the cortex of the root leading to formation of nodules. The root nodues are the actual sites of biological nitrogen fixation by the enzyme nitrogenase. In root nodules the, the oxygen sensitivity of nitrogenase is taken care by special hemoglobin - leghemoglobin. The globin protein is encoded by plant genes but the heme cofactor is made by the symbiotic bacteria (Appleby, 1984).

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. The genus *Frankia* consists of filamentous actinomycetes forming symbiotic associations with a number of woody dicot plants like *Casuarina, Hippophae, Alnus, Myrica*, etc. belonging to different families (Benson and Silvester, 1993). *Frankia* compartmentalizes nitrogenase within the vesicle structures, which are surrounded by an envelope containing a high content of bacteriohopane lipids and function to protect the enzyme from oxygen inactivation (Berry *et al*. 1993; Huss-Danell, 1997). Over years diazotrophy have been reported from other actinomycetes as well such as *Mycobacterium flavum*, *Corynebacterium autotrophicum, Arthrobacte*r sp., *Agromyces etc.* (Gtari *et al*., 2012). The findings of several authors (Von Bulow and Dobereiner, 1975; Dobereiner, 1976; Baldani and Baldani, 2005) revealed existing associations of tropical grasses with nitrogen-fixing bacteria, which under favorable conditions, contribute significantly to the N economy of these plants. The bacteria belong to the genus *Azospirillum* and are the most promising microorganisms that colonize roots of economically important grasses and cereals (Leigh, 2002).

Cyanobacteria have long been known to fix nitrogen. Both heterocystous and non heterocystous cyanobacteria like *Anabaena, Nostoc, Trichodesmium, Lyngbya, Plectonema* are capable of diazotrophy (Schlegel and Zaborosch, 2003). They are the only organisms that are capable of both O_2-evolving photosynthesis and nitrogen fixation (Klipp, 2004). Therefore, face the unique problem of balancing two essential, but incompatible, cellular

processes: oxygenic photosynthesis and O_2-sensitive N_2 fixation. In some filamentous cyanobacteria, nitrogen fixation occurs in specialized, terminally differentiated cells called heterocysts that protect the nitrogenase complex from O_2 damage by increasing respiration, terminating photosystem II activity, and forming multi-layered cellular membranes that reduce oxygen diffusion thus creating a microaerobic environment (Adams, 2000). However, in members like *Lyngbya*, *Plectonema* etc. where heterocyst is absent nitrogen fixation occur in internally organized cells (Schlegel and Zaborosch, 2003). Another important aspect of cyanobacteria, is their association with higher plants. The *Anabaena-Azolla* association (Bohlool *et al*., 1992), *Nostoc–Gunnera* association (Mylona *et al*., 1995) can fix a substantial amount of nitrogen. Cycads in association with cyanobacterial species can also fix nitrogen (Rai *et al*., 2002)

2.3 Major features of nitrogen fixation research

2.3.1 Nitrogenase complex: structural features and biochemistry

The biochemical machinery required for biological nitrogen fixation is provided by the nitrogenase enzyme system (Eady and Postgate, 1974; Hoffman *et al*., 2009). Nitrogenase is a two protein component system that catalyzes the reduction of dinitrogen to ammonia coupled to the hydrolysis of ATP (Rees and Howard, 2000). The most extensively studied form of nitrogenase is the molybdenum-containing system that consists of two component metalloproteins, the molybdenum-iron (MoFe-) protein and the iron (Fe-) protein (as seen in Figure2.1) . The smaller component of nitrogenase is the Fe protein, which acts as a redox-active agent and transfers electrons to the MoFe protein for reduction of substrates from available electron donor in the system (Rees *et al*., 2005). It has two identical subunits. The Fe protein contains one iron sulfur cluster [4Fe-4S], which bridges the two subunits. The oxidation state of the Fe protein can vary through the iron sulfur cluster. The Fe protein has one MgATP binding site in each subunit that binds to two MgATP molecules. Binding of MgATP to the Fe protein induces conformational changes followed by hydrolysis of MgATP, which facilitate the electron transfer from the Fe protein to the MoFe protein (Rees *et al*., 2005). Although this transfer of electrons is

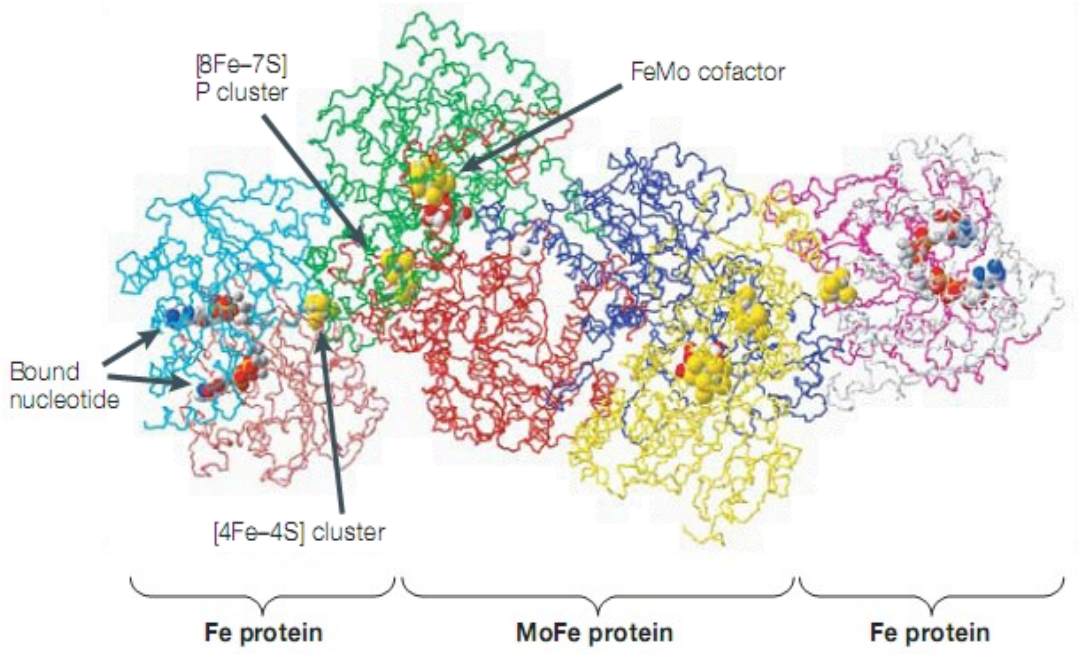


Figure 2.1 Structure of the complex that is formed between the Fe and MoFe protein component of *Azotobacter vinelandii* nitrogenase (Reproduced from Dixon and Kahn (2004) with author's permission)

the main function of the Fe protein, it has some other functions also. The Fe protein is needed for initial biosynthesis of the MoFe cofactor. Following the biosynthesis of MoFe cofactor, the insertion of the preformed MoFe cofactor into the MoFe protein requires the Fe protein (Burgess and Lowe, 1996). The larger component of nitrogenase is the MoFe protein, which is a $\alpha_2\beta_2$ tetramer, containing two αβ dimer subunits. Each dimer contains one MoFe cofactor and one P-cluster, [8Fe-7S]. The MoFe cofactor is located in the active site of the protein where the reduction of substrates occurs. The main role of the P-cluster is electron transfer by accepting an electron from the Fe protein and donating it to the MoFe cofactor. Each cluster contains eight metals and associated sulfurs that are distinctively arranged in ways that have been not been observed in any other enzymes The αβ dimeric units communicate and contact each other through their subunits (Burgess and Lowe, 1996). The P cluster bridges between each α and β subunit while the MoFe cofactor is placed on the α subunits. In addition to this molybdenum containing nitrogenase, alternative nitrogenases also exist that are homologous to this system, but with the molybdenum almost certainly substituted by vanadium or iron (Eady, 1996). The vanadium-nitrogenase

system has two components. It has a Fe protein which is the same as other nitrogenase systems and the second component is a vanadium-iron (VFe) containing protein which is different than two other systems. This type of nitrogenase has been detected in *A. vinelandii* and *A.chroococcum* (Robson *et al*., 1986). The third type of nitrogenase, iron only, contains an iron (Fe) protein and another protein, which is very similar to MoFe protein and VFe protein, but has only Fe as its cofactor. This type of protein has also been detected in *A. vinelandii* nitrogenase (Eady, 1996).

Studies (Thorneley and Lowe, 1985; Burgess and Lowe, 1996) revealed that the basic mechanism of nitrogenase involves: (i) complex formation between the reduced Fe-protein with two bound ATP and the MoFe-protein; (ii) electron transfer between the two proteins coupled to the hydrolysis of ATP; (iii) dissociation of the Fe-protein accompanied by re-reduction (via ferredoxins or flavodoxins) and exchange of ATP for ADP and (iv) repetition of this cycle until sufficient numbers of electrons and protons have been accumulated so that available substrates can be reduced. In addition to dinitrogen reduction, nitrogenase have been found to catalyze the reduction of protons to dihydrogen, as well as non-physiological substrates such as acetylene. Under optimal conditions, the overall stoichiometry of dinitrogen reduction is best described by the equation:

$N_2 + 8e^- + 8H^+ + 16MgATP \rightarrow 2NH_3 + H_2 + 16MgADP + 16P_i$

It has been shown that nitrogenase turnover requires an electron donor in addition to adenosine triphosphate (ATP). Electrons are generated in vivo either oxidatively or photosynthetically, depending on the organism. These electrons are transferred to flavodoxin or ferredoxin, a (4Fe-4S) -containing electron carrier that transfers an electron to the Fe protein of nitrogenase, beginning a series of oxido-reduction cycles. Electron transfer must occur six times per each fixed N2 molecule so that a total of 12 ATPs are required to fix one N2 molecule. However, nitrogenase also reduces protons to H2, a reaction which consumes two electrons. Therefore, the total cost of N2 reduction is eight electrons transferred and 16 MgATPs hydrolyzed. Hageman and Burris (1978) reported that each electron-transfer step requires an obligatory cycle of association of the

Fe and MoFe proteins to form a complex, after which the two components dissociate. Nitrogenase have also been found to be relatively slow enzyme with a turnover time of ~5 s^{-1}, and dissociation of the complex is the rate-limiting step (Dixon and Kahn, 2004).

Over the years, structural enzymology of nitrogenase has been the topic of many research projects. Following the initial structure determinations (Georgiadis *et al*., 1992; Kim and Rees, 1992) a collection of structures from different organisms and in different states are now available for the MoFe-protein in public domain (Bolin *et al*., 1993a; Kim *et al*. 1993; Peters *et al*. 1997; Mayer *et al*. 1999; Einsle *et al*. 2002); Fe-protein (Georgiadis *et al*. 1992; Schlessman *et al*. 1998; Jang *et al*. 2000; Strop *et al*., 2001; Sen *et al*.. 2004) and complexes between the two proteins (Schindelin *et al*., 1997; Chiu *et al*., 2001; Schmid *et al*., 2002). Crystallographic details of the structure of nitrogenase iron protein (Georgiadis *et al*., 1992) revealed that the protein dimer consisted of mixed α-helix/β-sheet polypeptide fold, with a consensus topology of an eight-stranded β-sheet flanked by nine α-helices. A β-sheet core in each monomer was formed of one short antiparallel and seven parallel β-strands. The binding sites critical to NifH function was determined by loops at the carboxy-terminal ends of several β-strands. Rees *et al.,* (2005) stated that Fe protein (NifH) adopts a polypeptide fold, commonly seen in P-loop containing nucleotide binding proteins such as ras p21, G-proteins, myosin and related proteins. The defining characteristic of members of this nucleotide-switch family is that they acquire different conformations depending upon the phosphorylation state of the bound nucleotide. The switch between conformations occurs with the chemical conversion of the triphosphate to the diphosphate form of the nucleotide, so that the rate of nucleotide turnover (the combination of hydrolysis and nucleotide exchange) on these proteins is the ultimate timing event for subsequent cellular processes. Structural comparisons (Lahiri *et al*., 2008) revealed that NifH is similar to the nucleotide binding proteins based on the presence of the structural features such as: (i) parallel β-sheets flanked by α-helices (ii) a phosphate-binding loop (P-loop) or Walker A motif, containing the G-X- X- X- X-G-KS/T consensus sequence and (iii) two

switch regions, Switch I and Switch II that interact with the γ- phosphate group of the bound nucleoside triphosphate. Studies have shown that the MoFe protein is an $\alpha_2\beta_2$ heterotetramer, where the α and β subunits exhibit similar polypeptide folds consisting of three domains of the α/β-type, with some extra helices (Rees and Howard, 2000). The individual domains share common structural elements, in particular, a central four stranded parallel β-sheet with flanking helices that is also found in the metallocluster binding domains of Fe-hydrogenase and CO-dehydrogenase (Rees, 2002). According to Rees, the α and β subunits within a dimer are approximately related by a twofold rotation axis. Each αβ dimer contains one FeMo-cofactor and one P-cluster, with the FeMo-cofactor binding in a cleft at the interface between the three domains of the α subunit, while the P-cluster is buried at the interface between a pair of α and β-subunits. The tetramer contains 30 Fe and two Mo atoms, which are distributed between two types of cluster, the P-cluster (8Fe8S) in the Rees model or (8Fe7S) in the Bolin model and the FeMo-cofactor (seven Fe, one Mo, nine S, one homocitrate) (Bolin *et al*. 1993b; Rees and Howard 2000).

Mutant studies have been instrumental in many systems for analyzing the contributions of specific residues to the catalytic mechanism of enzymes. In the case of nitrogenase, site directed mutagenesis studies have identified substitutions of residues around the FeMo-cofactor that have altered substrate reduction properties. Kim and his Co-workers (1995) reported about a mutant protein where His α195 residue had been replaced by Glutamine residue. His-195 residue is responsible for donating a hydrogen bond to one of the bridging sulphurs of the cofactor. Replacement of this histidine resulted in an enzyme that can still reduce acetylene to ethylene yet cannot reduce dinitrogen. This experiment showed that dinitrogen binding to the FeMo-cofactor is necessary, but not sufficient, for catalysis. In another experiment, Site-directed mutagenesis of the *Klebsiella pneumoniae* nitrogenase in the five conserved cysteine residues 63, 89, 155 and 275 present in the α- subunit and the three conserved cysteine residues 69, 94 and 152 present in the β-subunit of nitrogenase component 1 resulted in a loss of diazotrophic growth (Kent *et al*., 1989). Site directed mutagenesis

have also been reported from nitrogenase iron protein in *Azotobacter vinelandii* where alanine have been replaced by serine at amino acid 157 leading to disruption of nitrogen fixing ability of the microbe (Bursey and Burgess, 1998).

2.3.2 Genetics and Genomics of Biological N_2 fixation:

The biochemical complexity of nitrogen fixation is reflected in the genetic organization and in the regulation of expression of the components required for the catalytic activity. Various techniques like mutations, deletion mapping, cloning vectors etc. have facilitated the identification of genes associated with nitrogen fixation. The organization and regulation of the genes were revealed in the early 1980's. The organism that appears to have the simplest organization of nitrogen-fixation-specific (*nif*) genes, and which is the one best studied at the molecular genetic level, is the facultative anaerobe, *Klebsiella pneumoniae*. Arnold *et al*., (1988) reported the first ever detailed organization of *nif* genes from this organism. A 24 kb base pair DNA region, contains the entire *K. pneumonia nif* cluster, which includes 20 genes (as seen in Figure 2.2) . *nif*HDK are the three structural genes encoding for the three subunits of Mo nitrogenase. In most nitrogen fixing prokaryotes, these three genes form one transcriptional unit, with a promoter in front of the *nif*H gene. A number of studies (Dixon *et al*., 1980; Paul and Merrick, 1989; Rubio and Ludden 2005; 2008) have established that the maturation of apo-Fe protein (NifH) requires the products of *nif*H, *nif*M, *nif*U and *nif*S, while that of apo-MoFe protein requires at least six genes *nif*E, *nif*N, *nif*V, *nif*H, *nif*Q and *nif*B which are required for the biosynthesis of FeMoco. There is considerable homology between *nif*DK and *nif*EN, and it has been speculated that the *nif*EN products might form a scaffold for FeMoco biosynthesis that later shifts FeMoco to the *nif*DK complex (Brigle *et al*., 1987). Imperial and his co workers (1984) established that the *nif*Q gene product might be involved in the formation of a molybdenum-sulfur precursor to FeMoco and it has a typical motif characteristic of metal-binding sequences (Cys-X4-Cys-X2-Cys-X5-Cys) . Mutations in *nif*B result in the formation of an immature MoFe protein that lacks FeMo cofactor. It can be activated in vitro by adding FeMo

cofactor that has been isolated from wild type MoFe protein (Roberts *et al*., 1978). Mutations in the *nif*V gene result in the formation of a nitrogenase with a bound citrate rather than homocitrate. The *nif*V product is homocitrate synthetase (Zheng *et al*., 1997). Thus, on the basis of mutational studies, the function of various other *nif* genes has been confirmed. In contrast to *Klebsiella*, the *nif* organization is a bit complex in *Azotobacter vinelandii.* In, *Azotobacter* the genes coding for the Mo-dependent nitrogenase components (*nif*HDK) and their regulatory and assembly systems are located in two discrete regions (O'Carroll and Dos Santos, 2011). The organization of nitrogen fixing genes along with their genetic regulation in different rhizobia was extensively reviewed by Fischer (1994) and according to him Rhizobial *nif* genes are structurally homologous to the 20 *K.pneumoniae nif* genes, and it is inferred that a conserved *nif* gene plays a similar role in rhizobia as in *K. pneumoniae*. The organization of nitrogen fixation (*nif*) genes in cyanobacteria differs significantly from that found in other diazotrophs. In heterocystous cyanobacterium *Anabaena* sp, strain PCC 7120, the major *nif* cluster comprises of several

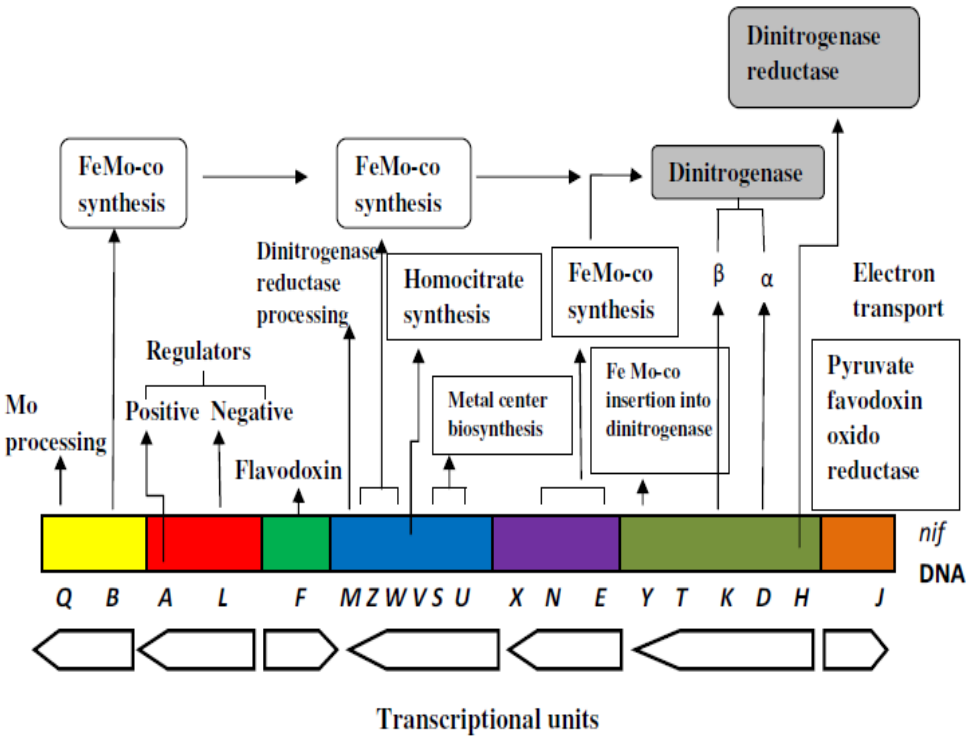


Figure 2.2: The nif gene cluster in *Klebsiella pneumoniae*, showing the arrangement of nif genes along with their functions.

open-reading frames (ORFs), two interrupting DNA elements and just four Operons dispersed over nearly 80 kb of DNA (Mazur *et al*., 1980; Rice *et al*., 1982; Golden *et al*., 1985). At least 14 genes associated with nitrogen fixation have been identified in *Anabaena* sp. strain PCC 7120. These include the *nif*K, D, H operon (Mazur *et al.,* 1980; Rice *et al*., 1982), the *nif*B, *fdx*N, *nif*S, *nif*U operon (Mulligan and Haselkorn, 1989), a *fdx*H gene which encodes a heterocyst-specific ferredoxim (Bohme and Haselkorn, 1988), and another operon containing *orfl*, *orf2* which result in a Nif $^-$ phenotype,if mutated (Borthakur *et al*., 1990). Study by Golden *et al*., (1985), established that the *nif*D and *fdx*N genes in *Anabaena* were discontinuously organized on account of two large interruptions and the interrupting DNA are excised by means of recombinases encoded by the *xis*A and *xis*F genes during heterocyst development.

Two of the *nif* genes, *nif*L and *nif*A, have been implicated to be the regulatory genes of the nitrogen fixation system (Arnold *et al*., 1988). According to Dixon and Kahn (2004), Nif A is a *nif*-specific transcriptional activator protein which interacts with the RNA polymerase σ-factor, σ54 (the product of *rpo*N) to initiate transcription at *nif* promoters. NifA belongs to a family of transcriptional activators all of which act in concert with an RNA polymerase holoenzyme containing the alternative σ54. The formation of open complexes and transcription initiation is catalyzed by NifA and involves nucleoside triphosphate hydrolysis (Morett and Buck, 1989). Austin and his co-workers (1994) were able to isolate NifA protein in its native form from *A. vinelandii*. The activity of NifA is usually regulated in response to two major environmental factors, dioxygen and fixed nitrogen. The NifL protein is an inhibitor of *nif* gene transcription in many systems. NifA activity is inhibited by NifL and requires stoichiometric amounts of the two proteins, implying that the mechanism of inhibition is by direct protein-protein interaction rather than by catalytic modification of the NifA protein (Lei *et al*., 1999) This interaction facilitates regulation of NifA activity in response to dioxygen, fixed nitrogen, and energy status, each of which is signaled by an independent pathway (Merrick *et al*., 2000).

Beside the *nif* genes, the “fix” and

"nod" type genes are associated with biological nitrogen fixation and nodule formation in rhizobial species, and many do not have homologues in the free–living diazotroph like *K. pneumonia*. The *fix* genes represent a very heterogeneous class including genes involved in the development and metabolism of bacteroids. Studies by Anthamatten and Hennecke, (1991); Batut *et al*., (1991) have established that *fix*L, *fix*J and *fix*K genes encode regulatory proteins. The *fix*ABCX genes code for an electron transport chain to nitrogenase (Fischer, 1994). Mutations in any one of the *fix*ABCX genes of *S. meliloti, B. japonicum*, and *A. caulinodans* completely abolish nitrogen fixation. All four *fix*GHIS gene products are predicted to be transmembrane proteins, but further biochemical analysis is required to define their function in rhizobial nitrogen fixation (Fischer, 1994). The *fix*NOQP genes encode the membrane-bound cytochrome oxidase that is required for respiration of the rhizobia in low-oxygen environments (Delgado *et al*., 1998). Johnston and his co-workers discovered the presence of nodulation genes in a plasmid of *Rhizobium leguminosarum* and mutation of those genes rendered them useless. Later on studies (Schultze 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). In most species, the *nod* ABC genes are part of a single operon. Inactivation of these genes abolishes the ability to elicit any symbiotic reaction in the plant (Long 1989). Over the years other *nod* genes like *nod*D, *nod*EF, *nod*S, *nod*L, *nod*HPQ have been characterized in many rhizobia. Like the rhizobia, *Azosprillium* includes a megaplasmid and sequences similar to *nod* genes (Elmerich, 1984). *Frankia* on the other hand houses a number of *nif* genes but researchers failed to spot *nod* genes in *Frankia* (Ceremonie *et al*., 1998).

Understanding of genetic machinery behind biological nitrogen fixation attained new heights with the arrival of complete genome sequences of various diazotrophs. Recent advances in genome sequencing have opened exciting new perspectives in the field of genomics by providing the complete

gene inventory of rhizobial microsymbionts. Genomics have enabled thorough analysis of the gene organization of nitrogen-fixing species, the identification of new genes involved in nitrogen fixation, and the identification of new diazotrophic species. *Mesorhizobium loti* strain MAFF303099 (Kaneko *et al*., 2000) was the first symbiotic bacterium to be sequenced 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* (Gonzalez *et al*., 2006), *Bradyrhizobium* strains, *Frankia* strains (Normand *et al*. 2007) and sequences for a number of free-living diazotrophs spanning different habitat and ecological niches bolstered nitrogen fixation. The genome information from all these nitrogen-fixing organisms allow researchers to rapidly apply information obtained from genome sequencing to the developing area of functional genomics, which will provide new insights into the complex molecular relationships that support both symbiotic and non-symbiotic nitrogen fixation. DNA array technologies are now being used to monitor the expression of a whole genome in a single experiment. The first massive approach to transcriptional analyses of entire symbiotic replicons was based on a high-resolution transcriptional analysis of the symbiotic plasmid of *Rhizobium* sp. NGR234 (Perret *et al*., 1999) at the Universite de Geneve, which developed methods to study the regulation of bacterial genes during symbiosis. The transcriptome for *S. meliloti* has been examined under a variety of conditions, including in planta (Ampe *et al*., 2003; Berges *et al*., 2003). Genome-wide macroarray experiments performed upon bacteroids from *M. loti* strain MAFF303099 resulted in the identification of expression islands scattered across the symbiotic island (Uchiumi *et al*., 2004). Proteome analysis has also been applied in nitrogen fixation research. Guerreiro *et al*.,(1999) established a total cellular proteome reference map of *S. meliloti* by comparing cells of the early exponential phase with the late exponential phase. Chen *et al*., (2000) detected 60 differences between the protein profiles of the *S. meliloti* strain and its pSymA-cured mutant derivative. The majority of these differences were due to regulatory

changes. Djordjevic and colleagues, using the symbiosis between *S. meliloti* and *Melilotus alba* (Natera *et al*., 2000), characterized novel symbiosis proteins and determined how the two symbiotic partners alter their respective metabolisms as part of the interaction. Over the years, geneticists have shifted their focus towards the macrosymbiont plant partner and have been involved in identification of the network of plant genes which initiate root nodule formation and support the nitrogen fixation process. Expressed sequence tag (EST) sequencing program are of particular relevance for the identification of genes involved in nodule formation. In this respect, *Glycine max* and *M. truncatula* proved to be particularly valuable in terms of EST sequences available in the GenBank dbEST division. Continuing initial efforts in *M. truncatula* by Covitz *et al*., (1998) and Gyorgyey *et al*., (2000), more than 170000 *M. truncatula* ESTs were deposited in dbEST. High-throughput EST analyses by Mathesius *et al*., (2001); Colebatch *et al*., (2002) in various legumes have been quite fruitful in identification of upregulated genes of root nodules. Recent completion of the whole genome of model legume *Medicago trunculata* (Young *et al*., 2012) can prove to be an important milestone for identification of genes that control the root nodule organogenesis.

2.4 Emergence of interdisciplinary field of bioinformatics:

It was the year, 1953 when James Watson and Francis Crick proposed the twisted-ladder (double helix) structure of deoxyribonucleic acid (DNA). This particular event forever changed the history of biological science and gave rise to modern molecular biology. In short order, their discovery yielded ground-breaking insights into the genetic code and protein synthesis. During the 1970s and 1980s, it helped to produce new and powerful scientific techniques, specifically recombinant DNA research, genetic engineering, rapid gene sequencing etc. Around the same time, the term “bioinformatics” was coined by Paulien Hogeweg and Ben Hesper (Hogeweg, 2011). Bioinformatics is by nature a cross-disciplinary field that began in the 1970s with the efforts of Margaret O. Dayhoff, Walter M. Fitch, Russell F. Doolittle and others and has gradually matured into a fully developed discipline. Initially, it was referred as "the study of information processes in biotic systems"(Hesper and Hogeweg,

1970). However, its primary use since at least the late 1980s has been to describe the application of computer science and information sciences to the analysis of biological data, particularly in those areas of genomics involving large-scale DNA sequencing (Luscombe *et al*., 2001). The arrival of the INTERNET is another important milestone in the development of bioinformatics as a full-fledged discipline. This discipline represents the convergence of genomics, biotechnology and information technology, and encompasses analysis and interpretation of data, modeling of biological phenomena, and development of algorithms and statistics (Fenstermacher, 2005). The need for bioinformatics was further accelerated when the human genome project (HGP) was launched in 1989. The aim of the project was to sequence the entire human genome. The information gleaned from the HGP is not very useful until the huge data is managed and interpreted in a proper way by the computational tools leading to the materialization of bioinformatics. The success of human genome project opened the flood-gates for other genome sequencing projects. Gradually genome sequences of mouse, rat, worms, yeast and plants like rice, *Arabidopsis* were completed. Currently there are more than Four thousand bacterial genomes in public domain. The publication of huge amount of sequence data were greatly supported by development of high end computers, smart computational tools for large-scale annotation, functional classification of the proteins (Searls, 2000) and development of specific databases (Birney *et al*., 2002). Availability of complete genome sequences for different organisms lead to the development of "public" repositories of gene data like GenBank (Benson *et al*., 2000), EMBL (Baker *et al*., 2000), DDBJ (Okayama *et al*., 1998) or the Protein DataBank (the PDB) (Bernstein *et al*., 1977). After the formation of the databases, tools became available to search sequence databases. Two programs, which greatly facilitated the similarity search, were FASTA (Pearson and Lipman, 1988) and BLAST (Altschul *et al.,* 1990). Many programs have been further developed from them. Accessibility of free and open source software has taken bioinformatics and its application to all-together new heights.

2.4.1 Application of Bioinformatics in

BNF research:

As we enter into the post genomics era, the bioinformatics tools have emerged as important means in research of biological nitrogen fixation. Large-scale genome 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 proteins, codon usage etc. With the current deluge of data, computational methods have become indispensable to biological investigations. The development of bioinformatics and statistical genetics, have resulted in the production of a number of tools, which are used to annotate the genome and obtain productive information from them (Hogeweg, 2011). Originally developed for the analysis of biological sequences, bioinformatics now encompasses a wide range of subject areas including structural biology, genomics and gene expression studies.

One of the primary applications of bioinformatics is the organization of the biological data in database that allows researchers to access existing information with ease. Open Access databases like GenBank, EMBL, DDBJ now houses thousands of *nif*H and *nif*D sequences. The numbers of fully sequenced and assembled diazotrophic genomes deposited in the databases have also gone up in the last few years. Simultaneously, new databases exclusively devoted to various aspects of biological nitrogen fixation like NodMutDB (Nodulation Mutant Database) (Mao *et al*., 2005), RhizoGATE (Becker *et al*., 2009), RhizoBase (http://genome.kazusa.or.jp/rhizobase/) etc has also surfaced in recent years. EST program conducted in the model legume *M. truncatula*, has led to the development of databases that allow data mining to identify genes relevant for nitrogen-fixing symbioses, for example, the TIGR M. truncatula Gene Index (http://www.tigr.org/tdb/mtgi) (Quackenbush *et al*., 2000), the *M. truncatula* database MtDB2 (http://www. medicago.org) and the database of the *Medicago* Genome Initiative (Bell *et al*., 2001). The data present in the various databases can be analyzed and interpreted in a biologically meaningful manner with the aid of computational tools.

Nowadays, the rapid increase in the number of prokaryotic species with sequenced genomes enables the development of *in-silico* searching tools to identify complex biochemical pathways such as nitrogen fixation.

Such assumptions, although very accurate, yield putative results and do not obviate the need for genetic and biochemical confirmation of gene function. Computation prediction tools like BLAST (Basic Local Alignment Search Tool) are being used by researchers for examining the occurrence and distribution of nitrogen fixation genes. The genomes present in the database are being scanned using NifHDK as query sequence (O'Carroll and Dos Santos, 2011). Phylogenies for the major *nif* operon genes have been inferred by distance matrix based methods like neighbour-joining or UPGMA or Maximum-likelihood based methods in attempt to understand the timing and complex genetic events that have marked the history of nitrogen fixation (Raymond *et al*., 2004). Computational tools are also now routinely employed by researchers (Amadou *et al*., 2008; Carvalho *et al*., 2010; Peralta *et al*., 2011; Black *et al*., 2012) to compare entire genomes of dizotrophs, which permits the study of more complex evolutionary events, such as gene duplication, horizontal gene transfer, and the prediction of factors important in bacterial speciation. Comparative genomics of *Frankia* yielded vital information regarding their evolutionary history and linked the inconsistency of genome size with the biogeographic history of the host plants harboring the microbial strains (Normand *et al*., 2007). Systems biology is another area where computer based simulation has been used extensively to analyze and visualize the complex connections and circuits of cellular pathways such as nitrogen fixation. Zhao and his collegues (2012) used several *in-silico* tools for reconstruction of metabolic network involved in symbiotic nitrogen fixation in *S. meliloti* 1021. It provided a knowledge-based framework for better understanding the symbiotic relationship between rhizobia and legumes. Bioinformatics is also indispensable for the examination of the data obtained in proteome analysis. An excellent resource of Internet-accessible proteome databases is the Expert Protein Analysis System (ExPASy), available online at http:// www.expasy.ch/ (Gasteiger *et al*., 2003). Furthermore software packages have been developed that can take multiple protein-expression profiles and automatically identify quantitative changes of interest. Two-dimensional electrophoresis databases are accessible on the Internet and can be

browsed with interactive software and integrated with in-house results. A cluster of Orthologous Groups of proteins (COG) is a new database search and represents an attempt at a phylogenetic classification of proteins from complete genomes (http:// www.ncbi.nlm.nih.gov/COG) (Tatusov *et al*., 2000). It serve as a platform for functional annotation of newly sequenced genomes and for studies on genome evolution. In addition, the identification of domains as subsets of proteins has been a very promising approach, implemented by databases such as InterPro (http://www.ebi.ac.uk/ interpro/). Proteomic analysis have revealed the direct genome functionality in a number of diazotrophic genomes (MacLean *et al*., 2007). Rapid developments of technological expertise in proteomics coupled with the improvement of *in-silico* tools have resulted in a deluge of structural information that guarantees acceleration in nitrogen fixation research.

As we march into the new millennium, practical application of computation tools to decipher meaningful information from available data is inevitable. Bioinformatics has the potential to elevate the research on biological nitrogen fixing bacteria and its protein machinery to a next level. Availability of bioinformatics tools have provided an opportunity to focus on the comparative genomics, molecular evolution of the genomes along with conformational and structural details of the proteins involved. Structural studies of proteins will provide a better understanding of the functional evolution of diazotrophy.

2.4.2 Research trends in codon usage analysis and comparative genomics:

In the post genomic era, application of bioinformatics tools in comparative genomics has led to the belief that every genome has its own story. Particularly the genetic code and its usage preferences is one of the most interesting aspects of biological science. Through the years several indices such as GC content, GC3 content, Fop (frequency of optimal codons) (Ikemura, 1985), relative synonymous codon usage (RSCU) (Sharp *et al*., 1986), effective number of codons (Nc) (Wright, 1990), Codon Bias Index (CBI) etc have been developed which proved to be significant in studies concerning codon usage patterns. 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. In the early period, majority of work on codon usage patterns focused upon *E. coli* (Peden, 1999). Gradually the bioinformatics analysis of codon usage was applied upon mammalian, bacterial, bacteriophage, viral and mitochondrial genes. Very preliminary work on codon usage of nitrogen fixing diazotrophs was initiated by (Mathur and Tuli, 1991). Ramseier and Gottfert (1991) reported differences in codon usage and GC content in *Bradyrhizobium* genes. Moderate codon bias was attributed to translational selection in nitrogen fixing genes of *Bradyrhizobium japonicum* USDA 110 (Sur, 2005). The analysis of synonymous codon usage patterns of three *Frankia* genomes (strains CcI3, ACN14a and EAN1pec) revealed that codon usage was highly biased, but variations were noticed among the three strains (Sen *et al*., 2008). Using codon adaptation index (CAI) highly expressed genes in *Frankia* were predicted. Synonymous codon usage analysis in *Azotobacter vinelandii* divulged considerable amount of heterogeneity (Sur *et al*., 2008). About 503 potentially highly expressed genes were identified and most of them were linked to metabolic functions of which 10 were associated with the core nitrogen fixing mechanism. Sen *et al*., (2012) explored the role of rare TTA codon in the genome of diazotrophic actinomycete *Frankia.*

Other than codon usage, molecular evolution of genes is another aspect which needs to be investigated. A more reliable index of genetic drift over evolutionary time is the ratio of Ka (nonsynonymous substitutions per site) to Ks (synonymous substitutions per site) for a large set of genes, based on comparisons of related species. The Ka/Ks ratio, which is almost always less than one, is widely used as an indicator of the extent of purifying selection acting to conserve coding sequences. This parameter has been widely applied in the analysis of adaptive molecular evolution, and is regarded as a general method of measuring the rate of sequence evolution in biology. These parameters have been used to assess the molecular evolution in plant hemoglobin genes (Guldner *et al*., 2004); secretory protein genes in *Streptomyces* and Yeast (Li *et al*., 2009) and in various

disease causing genes. Among diazotrophs, Crossman *et al*., (2008) measured the rates of synonymous (Ks) and non-synonymous substitutions (Ka) in orthologous genes of *R. etli* and *R. Leguminosarum.* More recently, synonymous and nonsynonymous substitution rates of orthologs shared by five species of the order Rhizobiales, three plant symbionts, one plant pathogen and one animal pathogen have been calculated by Peralta *et al*., (2011). Apart from the whole genome, molecular evolution of the genes responsible for symbiotic association and nodulation such as nodule specific genes (Yi, 2009) and recently *SymRK* (Mahe *et al*., 2011) have been specifically analyzed. But still a lot of symbiotic genes from wide range of diazotrophs have still to be analyzed to gather a complete scenario of their evolutionary rate in terms of their sequence features. Accumulations of bacterial whole genome sequences also give the biologists more opportunities to explore and compare the genomes in larger scale. Comparative genomics has given rise to a new concept highlighting the great diversity between closely related strains. A species can be described by its pangenome, i.e. the sum of a core genome containing genes present in all strains, and a dispensable genome, with genes absent from one or more strains and genes unique to each strain (Medini *et al*., 2005). Studying the diversity within pan-genomes is of interest for the characterization of the species or genus. Low pan-genome diversity could be reflective of a stable environment, while bacterial species with substantial abilities to adapt to various environments would be expected to have high pan-genome diversity (Snipen and Ussery, 2010). In 2005, Tettelin and colleagues introduced the conception of "pan-genome" in *Streptococcus agalactiae* (Tettelin *et al*., 2005). Soon afterwards, pan-genome has been widely used to provide insight into the analysis of the evolution of *S. pneumonia* (Hiller *et al*., 2007), *H. influenza* (Hogg *et al*., 2007), *E. coli* (Rasko *et al*., 2008), and so on. Besides evolution, pan-genome has been widely used to detect strain-specific virulence factors for some pathogens like *L. pneumophila* (D'Auria *et al*., 2010). Recently symbiotic pangenome of the nitrogen-fixing bacterium *Sinorhizobium meliloti* have been explored using computational methods and a set of accessory genetic factors related to the

symbiotic process have been defined (Galardini *et al*., 2011). As complete nucleotide sequences of more chromosome and symbiotic plasmids of nitrogen-fixing organisms become available, we have entered into the phase of comparative genomics. Comparative genomics also enables a much deeper understanding of the origin and evolution of free-living and symbiotic nitrogen fixation. Comparative genomics approach has been utilized by Carvalho *et al*., (2010) to delineate the evolutionary characterization of diazotrophic and pathogenic bacteria of the order Rhizobiales. Black *et al*., (2012) have worked upon fourteen strains of Rhizobiales to investigate the feasibility of defining a core 'symbiome'. Comparison of these whole genomes has revealed valuable information, such as several events of lateral gene transfer, particularly in the symbiotic plasmids and genomic islands that have contributed to a better understanding of the evolution of contrasting symbioses.

2.4.3 Bioinformatics of proteins related to BNF:

With recent developments in the field of proteomics, a large amount of biological data is now available in the public domain. This data include amino -acid sequences of nitrogenase proteins from a wide range of microbes. However, very little is known about the structure and role of all these proteins. Two technologies, X-ray and NMR, are by far the two most common means used to determine protein structure experimentally. In 1992, (Kim and Rees, 1992) provided a detailed crystallographic structure of molybdenum-iron protein of the *Azotobacter vinelandii* nitrogenase. The crystal structure of nitrogenase molybdenum iron protein has also been described from *Clostridium pasteurianum* (Kim *et al*., 1993). The X-ray crystal structure of *Klebsiella pneumoniae* nitrogenase component 1 (Kp1) has also been determined and refined to a resolution of 1.6 Å (Mayer *et al*., 1999). The 2.9 Å crystal structure of the NifH protein from *Azotobacter vinelandii* was obtained by (Georgiadis *et al*., 1992). However, tertiary structures of large number of nitrogenase proteins from different diazotrophs particularly those of symbiotic ones has not been yet resolved. The exact mechanism of working of these proteins is also relatively unknown due to the difficulty in obtaining crystals of

nitrogen bound to nitrogenase. This is because the resting state of MoFe protein does not bind nitrogen. Moreover, in the recent years quite a number of discrepancies have also crept out regarding the protein structures resolved by X-ray crystallography leading to retraction of papers (Chang *et al*., 2006). In this regard, a viable alternative approach is to predict 3D structure of proteins based on homology modeling technique and validate it properly. Homology modelling is a reliable technique that can consistently predict the 3D structure of a protein with precision akin to one obtained at low-resolution by experimental means (Marti-Renom *et al*., 2000). This technique depends upon the alignment of a protein sequence of unknown structure (target) with that of a homologue of known structure (template). This technique is particularly quite important in organisms with slow growth rate which poses difficulties in purification of subsequent proteins. Browne *et al.,* (1969) published the first report on homology modelling. A model of α-lactalbumin was constructed by taking the coordinates of a hen's egg-white lysozyme and modifying, by hand, those amino acids that did not match the structure. Since the mid 1980s, a large number of homology models of proteins with different folds and functions have been reported in the literature (Johnson *et al*., 1994; Sali, 1995). Homology modeling approaches was first applied for structural analysis of nitrogenase iron protein from *Trichodesmium* sp., a marine filamentous nitrogen-fixing cyanobacteria (Zehr *et al*., 1997). Standard homology modeling approaches have been also been used to generate reliable models of the nitrogenase Fe protein from thermophilic *Methanobacter thermoautotrophicus* based on structure of the *Azotobacter vinelandii* nitrogenase Fe protein (Sen and Peters, 2006) In recent times, the structure of the *Frankia* NifH protein was determined using homology modelling technique (Sen *et al*., 2010), Metal binding sites and functionally important regions of the protein were analyzed. These models based on homology are quite useful in providing conformational properties and structure -function relationship of these proteins.

A number of aspects of nitrogenase, particularly structure–function relationships are interesting areas of

fundamental research. The three dimensional structure of protein like that of nitrogenase is often considered an ideal model system for the study of the complex metal cluster mediated catalysis, electron transfer, complex metal cluster assembly, protein–protein interactions, and nucleotide dependent signal transduction. Molecular Dynamics simulations offers details about molecular motions as a function of time and are widely used to study protein motions at the atomic level. First protein simulation for 9.2 ps was carried out by McCammon *et al*., (1977) for bovine pancreatic tyrpsin inhibitor (BPTI). Case and Karplus (1979) work on dynamics of ligand binding to heme protein in 1979 is arguably the first simulation of ligand moving through the protein. First application of normal modes to identify low frequency oscillations using the energy minimization of the molecular mechanics force-field of protein was described by Brooks and Karplus (1983). This is the basic technique to identify domain-level motions in a protein. First simulation of a protein in explicit waters was done by Levitt and Sharon (1988).

Metalloproteins like nitrogenase are a vast class of biological molecules, which are responsible for many vital functions. Despite the intrinsic difficulties of these systems particularly those related to parameterization of the metal cofactors, they have been the object of several MD simulations. These studies are mainly focused on structural aspects, since the cluster has either a storage role or is involved in an electron-transfer process in these proteins. Amongst the metalloprotein having FeS cluster cofactor, molecular dynamics simulation has been carried for proteins like heme containing cytochrome P450 (Kuhn *et al*., 2001), Rubredoxins (Grottesi *et al*., 2002), 3Fe–4S cluster-containing protein, ferredoxin I (Meuwly and Karplus, 2004), Adenosine phosphosulphate reductase (dos Santos *et al*., 2009), hydA1 hydrogenase (Sundaram *et al*., 2010). More recently, molecular modeling, dynamics and docking studies on both *A. vinelandii* and *G. diazotrophicus* FeSII proteins and nitrogenases were carried out by Lery *et al.,* (2010), elucidating molecular aspects of protein-protein interaction. In the MD simulation of metalloproteins, the force field parameters of the metal ion and its ligands need to be defined beforehand

taking into account the nature of the metal ion, its coordination number, geometry, oxidation and spin states and the nature of its ligands. Several sets of parameters have been reported in the literature for the active sites of the most widely studied metalloproteins and including the coordination geometries of the metal ligand (Banci and Comba, 1997; Norrby and Brandt, 2001; Comba and Remenyi, 2002). One of the parameters that significantly affect the overall protein structure is the partial charges of the atoms of the metal–ligand moiety. In the bonded model, partial charges are commonly calculated through the RESP (Restrained Electrostatic Surface Potential) methodology (Fox and Kollman, 1998) based on semi-empirical or *ab initio* calculations. The *ab initio* calculations are mostly performed through density functional theory (DFT) calculations, with the B3LYP functional or Hartree Fock calculations (Banci, 2003). Thus, the development of proper parameters of the metal cofactors needs the amalgamation of quantum calculations in conjunction with classical molecular mechanics calculations. This will enable the description of not only structural features but also of reactivity properties of the metalloproteins.

Apart from molecular dynamics studies, availability of 3D structures has lead to the structural divergence studies of these proteins. The studies revealed that various proteins have structural and mechanistic similarities as well as evolutionary relationships with the nitrogenase iron protein, notable among them being: light independent protochlorophyllide (Pchlide) reductase (ChlL/FrxC or bChL), arsenite pump ATPase (ArsA), 2-hydroxyglutaryl dehydratase Component A (CompA) involved in glutamate degradation and MinD that functions in spatial regulation of cell division (Fujita *et al*., 1989; Fujita and Bauer, 2000; Gatti *et al*., 2000; Hu *et al*., 2003). The structures of Nif H, D, K homologs have been utilized for phylogenetic reconstruction based on a structural dissimilarity (RMSD) matrix (Boyd *et al*., 2011). Further, the investigation into the conservation of residues that line the active site pocket revealed that ancestor of all nitrogenases had an open cavity capable of binding metal clusters which conferred reactivity (Boyd *et al*., 2011).

2.4.4 Bioinformatics of BNF evolution:

2.4.4.1 Classical approach

Researchers have long sought to answer the question of when nitrogen fixation began and what evolutionary pressures affected it (Postgate and Eady, 1988; Berman-Frank *et al*., 2003). The emergence and evolution of nitrogen fixation ability (diazotrophy) among prokaryotes is complex and has not yet been fully elucidated. The incomplete distribution pattern of this highly conserved enzyme among Bacteria and Archaea has led to the development of conflicting hypotheses on BNF. The first idea theorizes that nitrogen-fixation is an ancient function of the last common ancestor of Bacteria–Archaea that was vertically transmitted, but has undergone widespread gene loss among descendants with horizontal transfer in some isolated instances (Hennecke *et al*., 1985; Normand and Bousquet, 1989; Fani *et al*., 2000; Berman-Frank *et al*., 2003). During this postulated time period, reduced nitrogen may have been very abundant and the initial function of nitrogenase was probably very different. One proposed initial function of ancient nitrogenase might be associated with detoxification mechanism for cyanides and other chemicals (Silver and Postgate, 1973; Fani *et al*., 2000). This idea is based on the observation that nitrogenase reduces a number of alternative substrates in addition to N2, several of which are toxins (e.g. cyanides) . The second hypothesis proposes that nitrogen fixation was an anaerobic ability that appeared after the emergence of oxygenic-photosynthesis and was subsequently lost in most lineages through horizontal transfer (Postgate, 1982; Postgate and Eady, 1988). Recently, Hartmann and Barnum (2010) examined Mo-nitrogenase phylogeny and proclaimed a conclusion combining both theories on diazotrophic evolution.

Nitrogenase genes are highly conserved at both the chemical and genetic levels across wide phylogenetic ranges and among closely related organisms. The conservation of nitrogenase genes lends itself for use as a genetic marker for phylogenetic analysis to help answer questions of the evolution of nitrogen fixation and its genes. 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. *nif*D and *nif*K are thought to be the result of an in-tandem gene

duplication (Fani *et al*., 2000; Postgate and Eady, 1988), giving the functional components of the enzyme. A second duplication event is thought to have occurred for the *nif*EN genes. Till date most of the studies concerning the evolution of nitrogen fixation have focused on the *nif* genes, primarily the highly conserved *nif*H gene but also the larger but less conserved *nif*D, *nif*K, *nif*E, and *nif*N genes (Normand and Bousque, 1989; Normand *et al*., 1992; Hirsch *et al*., 1995; Fani *et al*., 2000). Sequence alignment based methods are widely used to study the evolution of relevant *nif* genes. Young (2005) discussed the phylogeny and evolution of nitrogenases in details. According to Young, true NifH proteins can be divided into three types - **Type B** ("bacterial") is the best represented, and includes enzymes from the proteobacteria, cyanobacteria and firmicutes; Type **C** ("clostridial") is found in the firmicute bacterium and *Clostridium*, the green sulfur bacterium *Chlorobium*, and also in the archaeon *Methanosarcina* and **Type A** is associated with the "alternative" nitrogenases that do not contain molybdenum, and is found in both Archaea and Proteobacteria. There are also there are also a large number of more distant relatives, notable among them light independent protochlorophyllide (Pchlide). The similarity between these proteins and NifH was analysed and discussed by Burke *et al*., (1993), who argued that nitrogen fixation probably originated before photosynthesis, so the photosynthesis enzymes would have been derived from NifH rather than the other way round. The phylogenies of NifDKEN family have also been topic of many research works. Dedysh *et al*., (2004) utilized the NifD phylogeny to assesses the nitrogen fixation capabilities of methanotrophic bacteria. Henson *et al*. (2004b) 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. The strict requirement of NifH in biological nitrogen fixation and its universal presence in diazotrophs has resulted in this protein serving as a sequence tag or barcode for the identification of nitrogen fixers. Genomic analysis using the sequence of NifH as a query results in BLAST hits that include NifH, VnfH, and AnfH components of the Mo-, V-, and Fe-only nitrogenases, respectively (Raymond *et al*., 2004).

Recently, Dos Santos and his colleagues proposed a new criterion for computational prediction of nitrogen fixation: the presence of a minimum set of six genes coding for structural and biosynthetic components, namely NifHDK and NifENB (Dos Santos *et al*. 2012). Latysheva *et al.,* (2012) considered the various *nif* orthologs for performing empirical Bayes ancestral state reconstructions to investigate the evolution of nitrogen fixation in cyanobacteria.

Over the years, there has been a debate among the workers regarding horizontal gene transfer versus vertical descent as the dominant force in the evolution and distribution of N-fixation. In the case of an early origin and subsequent vertical descent of the *nif* genes, a comparison of SSU ribosomal phylogeny and the phylogeny of *nif* genes should reveal roughly the same features, assuming the mutation rates in both genes were similar. In the case of a late development and a mainly horizontal distribution of the genes, the phylogeny suggested by the *nif* genes should deviate significantly from the rRNA based standard tree. A number of researchers have presented strong evidence that SSU rRNA phylogeny and phylogeny based on the *nif* genes are in general agreement, suggesting that they have evolved in a similar fashion (Hennecke *et al*., 1985; Young, 1992; Zehr *et al*., 1997b). However, numerous studies have also highlighted instances of possible horizontal gene transfer in *nif*D (Parker *et al*., 2002; Qian *et al*., 2003; Henson *et al*., 2004a; Henson *et al*., 2004b), *nif*H (Normand and Bousquet, 1989; Hurek *et al*., 1997; Cantera *et al*., 2004; Dedysh *et al*., 2004) and *nif*K (Kessler *et al*., 1997) based on incongruence with 16S rRNA trees. Other studies have found support for both vertical descent and horizontal transfer (Hirsch *et al*., 1995). Haukka *et al*., (1998) proposed that horizontal gene transfer may have played an increasing role at genus and lower taxonomic levels. This may be especially important in organisms that have *nif* genes located on plasmids (Normand and Bousquet, 1989).

The phylogenies of alternative nitrogenases have also been under much speculation. Some organisms contain genes that encode alternative nitrogenases, based on vanadium and iron, or on iron alone that are expressed when Mo is unavailable (Eady, 1996). The nitrogenase encoded by the *vnf*HDK genes is believed to contain

vanadium in place of molybdenum in the active site cofactor, whereas the nitrogenase encoded by the *anf*HDK genes to contain only Fe as the metal constituent of its active site cofactor (Hales *et al*., 1986; Chisnell *et al*., 1988). Anbar and Knoll (2002) considered the conditions of prehistoric ocean and proposed that Anf and Vnf represent primitive forms of nitrogenase that predate Nif. It has been hypothesized that the sequences of the "alternative" cluster represent a different gene family that arose by gene duplication in a common ancestor of Bacteria and Archaea (Young, 1992). The nitrogenase systems (*vnf* and *anf*) found in some bacterial lines are highly related to the nitrogen fixing systems found in the archaeal *Methanosarcina* suggesting that lateral transfer may have occurred (Raymond *et al*., 2004). Glazer and Kechris (2009) examined the structural features and phylogeny of the α subunits of NifD (MoFe subunit), VnfD (VFe subunit), and AnfD (FeFe subunit) sequences. The result obtained by them is consistent with a radiation of NifDs, VnfD and AnfD from a common ancestral sequence. Recently, Boyd *et al.,* (2011), examined concatenations of protein homologs of the structural components shared between all known nitrogenases. The results of their phylogenetic and structure-based examination indicate an evolutionary path whereby Mo-dependent nitrogenase gave rise to the alternative forms suggesting that they arose later, perhaps in response to local Mo limitation.

2.4.4.2 Alternative approaches:

For tracing evolution of proteins within a set of divergently evolved proteins it is useful to construct the phylogenetic trees based on the similarities in the amino acid sequences and the base sequences of the genes. But previous studies seems to suggest that the origin and extant distribution of nitrogen fixation is perplexing from a phylogenetic perspective, largely because of factors that confound molecular phylogeny such as sequence divergence, paralogy, and horizontal gene transfer (Raymond *et al*., 2004). This lead to the assumption that sequence based phylogeny is not enough to reveal the complex evolutionary path in BNF. Moreover, many workers (Nadler, 1995; Qi *et al*., 2004; Sims *et al*., 2009) have also pointed out fallacies in sequence alignment based methods. Therefore alternate phylogenetic approaches are

being sought. Alignment-free condensed matrix method relying on nucleotide triplet is one such alternative approach. 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 approach has been utilized by researchers in phylogenetic analysis of Aminoacyl t-RNA Synthetase (Mondal *et al*., 2008), swine flu genomes (Sur *et al*., 2010), bacterial zeta toxin (Mondal *et al*., 2011) and nitrogenase proteins (Sur *et al*., 2010). Another suitable alternative of protein sequence alignments is the structure based phylogeny. It is well known that the 3-D structures and structural features of homologous proteins are conserved better than their amino acid sequences (Chothia *et al*., 1986; Hubbard and Blundell, 1987). It has been demonstrated several times that the homologous proteins could diverge beyond recognition at the level of their amino acid sequences but maintain similar structure and function. In several cases of low sequence similarity proteins retain the folds as well as broad biochemical features and functional properties, suggesting an evolutionary connection (Murzin *et al*., 1995; Russell and Sternberg, 1996). Hence, it is more appropriate to use similarities in 3-D structure of proteins rather than the amino acid sequence similarities in modelling evolution of distantly related proteins. Construction of phylogenetic trees using 3-D structures has been applied for a variety of protein families like Short-chain Alcohol Dehydrogenases (Breitling *et al*., 2001) and metallo-β-lactamases (Garau *et al*., 2005). Lately 3D structure-based phylogenetic approach have been utilized for functional characterization of proteins with cupin folds (Agarwal *et al*., 2009). It was revealed that structure-based clustering of members of cupin superfamily reflects a function-based clustering. Moreover, comparison of distance matrices, used in phylogenetic tree construction methods, has been considered as an equivalent of comparison of phylogenetic trees (Balaji and Srinivasan, 2001; Pazos and Valencia, 2001) to describe co-evolution of interacting partners and to study protein evolution. Therefore, structure based approaches can be utilized to assess the phylogenetic relationships of proteins involved in

BNF which shares low sequence similarity but high structural similarities with many proteins with diverse biological functions. Along with the trajectory of evolution of diazotrophy in various organisms another feature that needs attention is the functional divergence of the proteins involved in this biological process. Previous works (Gu, 1999; Dermitzakis and Clark, 2001; Raes and Van de Peer, 2003) have shown that gene duplication events often leads to a shift in protein function from an ancestral function resulting in functional divergence and as a consequence of which some residues are subjected to altered functional constraints. This implies that evolutionary rates at these sites will vary in different homologous genes of a gene family. Site-specific altered functional constraint (or shifted evolutionary rates) can be detected by comparing the rate correlation between gene clusters, when the phylogeny is given (Gu, 1999).This approach has been earlier exploited by researchers to trace the functional divergence in vertebrate haemoglobin (Gribaldo *et al*., 2003), G-protein alpha subunits (Zheng *et al*., 2007), OPR gene family in plants (Li *et al*., 2009), anoctamin family of membrane proteins (Milenkovic *et al*., 2010). But a broad picture on the functional divergence in the NifH/Bchl protein family is still unavailable.

Structural evolution of proteins is one of the last missing pieces of the puzzle called Biological nitrogen fixation. The concept of statistical coupling analysis (SCA) has been introduced for characterizing the pattern of evolutionary constraints on and between amino acid positions in a protein family. The traditional definition of conservation has been extended to include correlations between positions which will contribute to defining the architecture of functional interactions between amino acids, and ultimately, help understand the basic physical principles underlying protein structure, function, and evolution. Lockless and Ranganathan (1999) originally developed the notion of SCA. Statistical analysis of correlated evolution between amino acids have been utilized to identify groups of coevolving amino in S1A serine proteases (Halabi *et al*., 2009). Bachega *et al.,* (2009) used statistical coupling analysis for analyzing the metal binding specificity of SOD

enzymes of parasites. SCA has not been yet applied in the characterizing the proteins involved in BNF. This may provide vital clues for structural evolution of the proteins.

2.5 Challenges and future prospects:

Considerable progress has been made in understanding the machinery of biological nitrogen fixation in last decade. The major part of the research has been focused on the structure of nitrogenase, elucidation of the compositions and functions of all of the *nif*-gene products. In the past, problems associated with detection of *nif* gene from environmental samples and subsequent crystallization of the nitrogenase enzyme has been the major roadblocks in the BNF research. But as we enter into the post genomics era, the major hurdles have been removed. The challenge now is to put all the known information together and, with the combined application of biochemical, genetics and bioinformatics techniques, to determine how nitrogenase functions at the molecular level. With the rapid increase in the number of complete genomes of varied diazotrophs along with their nitrogen-fixing genes in the public domain, bioinformatics tools has emerged as potent weapon to tackle the unsolved mysteries of symbiotic and asymbiotic nitrogen fixation. It can be used to extract meaningful interpretation of sequence data. With the advent of *in-silico* protein modelling techniques and new algorithms for measuring structural divergence, the problems associated with functional evolution of nitrogenase system can also be tackled in a better way and new glimpses can be gained. Genomic studies aided by the bioinformatics tools 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.

Chapter 3
Materials and Methods

3.1 Codon and Amino acid Usage:

3.1.1 Retrieval of sequences:

Genomes sequences along with their amino acid sequences and annotations of various symbiotic and cyanobacterial diazotrophs along with free-living diazotrophs listed in Table 3.1, were acquired from the IMG database (www.img.jgi.doe.gov) (Markowitz *et al*., 2006). All protein coding genes, ribosomal protein genes and nitrogen fixation related genes were considered for codon usage variation analysis. The ribosomal protein genes are considered to be the most stable and highly expressed genes in a genome, hence they are considered along with nitrogen fixation related genes for the comparative analysis.

3.1.2 Codon Usage variation Analysis:

The codon usage disparity in various diazotrophic genomes have been investigated by several potent indices like amount of Guanine and Cytosine in the nucleotide sequence (G+C content), frequency of G and C in the third position of codons (GC3s), effective number of codons (Nc) (Wright, 1990), frequency of optimal codon (Fop) (Ikemura, 1985) and Codon bias index (CBI) (Bennetzen and Hall, 1982). All these parameters were calculated by the CodonW (Ver. 1.4.2) software (Peden, 1999). The parameters are crucial for determining the extent of codon usage bias in concerned genomes and highlight the factors affecting codon usage pattern.

The effective number of codons (Nc) is a parameter that can measure codon bias of synonymous codons. It is a quantitative measure reflecting the frequency of a small subset of codons used by a gene (Wright, 1990) and its value ranges from 20 (on usage of one codon per amino acid) to 61 (on usage of all the codons with equal frequency). Nc is computed as in equation (1)

$Nc=2+S+\{29/[S^2+(1-S)^2]\}$……....(1)

Where, *S* represents GC3 values, Frequency of optimal codon (Fop)

Table 3.1 List of diazotrophic genomes and proteomes considered for codon usage study

Group	Organism Name	Code	Accession#	Genome Size
Cyanobacteria	*Cyanobacterium* UCYN-A	CYAN	NC_013771	1.44
	Crocosphaera watsonii WH 8501	CROC	NZ_AADV02000001*	6.24
	Synechococcus sp. JA-3-3Ab	SYN-J3	NC_007775	2.93
	Synechococcus sp. JA-2-3B	SYN-J2	NC_007776	
	Nostoc punctiforme PCC 73102	NOSP	NC_010628	9.06
	Nostoc azollae 0708	NOSA	NC_014248	5.49
	Nostoc sp. PCC 7120	NOS7120	NC_003272	7.21
	Nodularia spumigena CCY9414	NOD	NZ_AAVW01000003*	5.32
	Anabaena variabilis ATCC 29413	ANA	NC_007413	7.11
	Lyngbya sp. PCC 8106	LYN	NZ_AAVU01000084*	7.04
	Cyanothece sp. PCC 7822	CY7822	NC_014501	7.84
	Cyanothece sp. ATCC 51142	CY51142	NC_010547	5.46
	Cyanothece sp. PCC 7424	CY7424	NC_011738	6.55
	Cyanothece sp. PCC 8801	CY8801	NC_011721	4.79
	Trichodesmium erythraeum IMS101	TRIC	NC_008312	7.75
Proteobacteria	*Rhizobium leguminosarum bv. viciae* 3841	RVICIA	NC_008379	7.75
	Rhizobium leguminosarum bv. trifolii WSM2304	RTRI	NC_011366	6.87
	Rhizobium etli CIAT 652	RCIAT	NC_010998	6.44
	Rhizobium etli CFN 42	RCFN	NC_007762	6.53
	Rhizobium sp. NGR234	RNGR	NC_000914	6.89
	Rhizobium rhizogenes K84	RRHIZ	NC_011983	7.27
	Bradyrhizobium japonicum USDA 110	BJAP	NC_004463	9.10
	Bradyrhizobium sp. BTAi1	BTAI	NC_009485	8.49
	Bradyrhizobium sp. ORS 278	BORS	NC_009445	7.46
	Mesorhizobium ciceri biovar biserrulae WSM1271	MCICER	NC_014918	6.69
	Mesorhizobium loti MAFF303099	MLOTI	NC_002678	7.6
	Mesorhizobium opportunistum WSM2075	MOP	NZ_ACZA01000080*	6.88
	Sinorhizobium medicae WSM419	SMED	NC_009621	6.82
	Sinorhizobium meliloti 1021	SMEL	NC_003037	6.69
	Azorhizobium caulinodans ORS 571	AZOR	NC_ 009937	5.37
	Azotobacter vinelandii DJ#	AZOTO	NC_012560	5.37
	Klebsiella pneumoniae 342	KLEB	NC_011283	5.76
	Beijerinckia indica indica ATCC 9039#	BEJ	NC_010581	4.42
Actinobacteria	*Frankia alni* ACN14a	ACN	NC_008278	7.5
	Frankia sp. CcI3	CCI3	NC_007777	5.43
	Frankia sp. EAN1pec	EAN	NC_009921	8.98
	Frankia symbiont of *Datisca glomerata*	DG	NZ_ADGT01000001*	5.34
	Frankia sp. EUN1f	EUN1f	NZ_ADGX01000001*	9.32
Green sulphur	*Chlorobium limicola* DSM 245	CHLO	NC_010803	2.76
	Chlorobium tepidum TLS	CHTE	NC_002932	2.15

('#' stands for free-living diazotroph ; '*' stands for Draft sequence)

represents the fraction of synonymous codons that are optimal codons (Ikemura, 1985). Its value ranges from 0 (meaning a gene has no optimal codons) to 1 (when a gene is exclusively comprised of optimal codons). Fop is generally determined by the equation 2(a):

$Fop = N_{oc}/N_{sc}$ …… (2a)

The original equation is modified in equation (2b), when rare codons are identified, as

$Fop_{(mod)} = N_{oc} - N_{rc} / N_{sc}$……..(2b)

Where N represents the frequency of each codon type used, whereas N_{oc}, N_{rc} and N_{sc} stands for optimal codons, rare codons and synonymous codons respectively.

The codon bias index (CBI) is another important estimator of directional bias and determines the level to which a gene uses a subset of optimal codons. The codon bias index value ranges from 0 to 1.It is calculated by equation (3) as follows:

$CBI = N_{opt} - N_{ran}/ N_{tot} - N_{ran}$……..(3)

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.

3.1.3 Prediction of expression pattern:

The codon adaptation index (CAI) is a commonly used measure of codon usage within a gene relative to a reference set of genes (usually ribosomal protein genes). The CAI values were computed using the web-based application the CAI Calculator2 (http://userpages.umbc.edu/~wug1/codon/cai/cais.php) (Wu *et al*., 2005). The CAI value varies between 0 and 1·0 (Sharp and Li, 1987), with higher CAI values indicating that the gene of interest has a codon usage pattern more similar to that in the reference genes. CAI is usually calculated based on the equation (4):

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

Where, ω_K signifies the relative adaptedness of the K^{th} codon and L represents the number of synonymous codons in the gene.

Ideally, the reference set in CAI is composed of highly expressed genes, so that CAI provides an indication of gene expression level under the assumption that there is translational selection to optimize gene sequences according to their expression levels. Generally ribosomal proteins are highly expressed in a given genome and as a consequence are taken as the reference set for calculation of CAI values. Statistical analysis including F-test of

variance, t-test and Pearson correlation was performed as described by Perriere and Thioulouse (2002).

3.1.4 Correlation of CAI and expression data:

The CAI values obtained for the protein coding genes of the *Frankia alni* ACN14a were correlated with the microarray expression data acquired for symbiotic and pure cultures (Alloisio *et al*., 2010). The data comprises fluorescence values in two different media i.e FrBAP+ (containing ammonium), FrBAP- (without $NH4^{+}$) and in symbiotic tissues of *Alnus glutinosa*.

3.1.5 Rose plot construction:

The difference of codon usage and amino acid usage for representative diazotrophic genomes is highlighted by the construction of Rose plot. It is a graphical representation of the frequency of codon usage along with amino acid usage of each of the bacterial genomes. The rose plot was constructed by a R-script described by Ussery *et al*., (2009).

3.1.6 Correspondence analysis of amino acid usage:

The amino-acid compositions of 40 diazotrophic genomes mentioned in Table 3.1 were calculated using the freqaa.pl script (http://www.pasteur.fr/~tekaia/HYG/scripts.html). The raw amino acid counts were transformed into relative compositions (or percentages). Based on the relative composition, a matrix of 40 observations (strains) and 20 variables (amino-acids) was obtained. Three supplementary variables were also considered in the analyses, these are : 'char' for charged amino acids (Asp (D), Glu (E), Lys (K), Arg (R) and His (H)), 'pol' for polar/uncharged amino acids (Gly (G), Ser(S), Thr (T), Asn (N), Gln (Q), Tyr (Y) and Cys (C)) and 'hyd' for hydrophobic amino-acids (Leu (L), Met (M), Ile (I), Val (V), Trp (W), Pro (P), Ala (A) and Phe (F)). The amino-acid composition values attributed to the supplementary variables were obtained by summing the respective contributions of the corresponding amino acids, in the various strains. To extract relevant information from amino acid compositional matrix, correspondence analysis was done with the help of XLSTAT program. 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). Thus, correspondence analysis helps in recognizing the diverse trends in the amino-acid compositions of the expressed proteins in various nitrogen fixing organism.

3.2 Nucleotide substitution pattern study:

3.2.1 Identification of orthologs:

The nucleotide sequences of all the protein coding genes of nitrogen-fixing bacteria enlisted in Table 3.2 were obtained from JGI-IMG Database (www.img.jgi.doe.gov) and considered for evolutionary rate analysis. Orthologous gene pairs between genome pairs were identified based on the Reciprocal Best Hits (RBH) approach on amino acid sequences for all the protein coding gene sequences with an E-value threshold of $1e^{-10}$; an identity $\geq$ 50 % and at least 50 % alignable region. This approach was used previously by Li *et al*., (2007) for screening orthologs in *Streptomyces.*

3.2.2 Evolutionary rate analysis:

The orthologous gene alignments were utilized for evolutionary rate analysis. The number of nonsynonymous or synonymous substitutions per site (Ka or Ks, respectively) and their ratio (Ka/ Ks) was estimated with Codeml in the PAML software program (Yang, 2007). A bioperl script was used for running the PAML to estimate the pairwise Ka/Ks values. The script first translates the cDNA into protein and then aligns the protein sequences. The protein alignment is then projected back into cDNA coordinates and finally provided to PAML module to calculate the Ka/Ks ratio using the Nei-Gojobori (NG) method (Nei and Gojobori, 1986). NG is a widely used unweighted pathway method of estimating synonymous substitutions.

3.2.3 Sliding window analysis of nifHDK gene:

Many times it is seen that different regions of a single gene can be exposed to different selective pressures. In these cases, calculating Ka/Ks over the entire length of the gene does not provide a clear picture of the evolutionary constraints associated with the gene. A sliding window method is usually adopted to examine the intragenic pattern of the substitution rates and to test for the occurrence of significant clusters of variant regions. A sliding window analysis was used for *nif*H, *nif*D and *nif*K genes to identify regions with differences in selective pressure due to functional constraints. JCoDA program (Steinway *et al*., 2010) was used for the regional pairwise Ka/Ks

calculation via sliding window. For sliding window analysis, a pair of sequences is selected from a list of all possible comparisons presented in the graphical user interface (GUI) of JCoDA. The size of the window, jump, and substitution models were also configured via drop down menu in the GUI. Once the selected sequence comparisons were submitted, JCoDA parsed through them and converted them to Phylip format which was then feed to yn00 program of the PAML package for Ka/Ks ratio calculation. The output in graphical format was then represented. yn00 is a part of the PAML package which is used for estimating synonymous and nonsynonymous substitution rates in pairwise comparison of protein-coding DNA sequences using maximum likelihood method

3.3 Comparative Genome Analysis:

3.3.1 Retrieval of genome data:

For comparative genome analysis, two completely different groups of symbiotic diazotrophs were chosen. The first group consists of 15 proteobacteria strains of the order Rhizobiales. The other group includes

Table 3.2 List of bacteria considered for evolutionary rate analysis

Genus	Strains	Code	Accession #
Bradyrhizobium	*Bradyrhizobium japonicum USDA 110*	BJAP	NC_004463
	Bradyrhizobium sp. BTAi1	BTAI	NC_009485
	Bradyrhizobium sp. ORS 278	BORS	NC_009445
Rhizobium	*Rhizobium etli CIAT 652*	RCIAT	NC_010998
	Rhizobium etli CFN 42	RCFN	NC_007762
	Rhizobium leguminosarum bv. trifolii WSM1325	RTRI	NC_012850
	Rhizobium leguminosarum bv. viciae 3841	RVICIA	NC_008380
Nostoc	*Nostoc sp. PCC 7120*	NOS7120	NC_003272
	Nostoc punctiforme PCC 73102	NOSP	NC_010628
	Nostoc azollae 0708	NOSA	NC_014248
Chlorobium	*Chlorobium phaeobacteroides*	CHPH	NC_008639
	Chlorobium limicola DSM 245	CHLO	NC_010803
	Chlorobium tepidum TLS	CHTE	NC_002932
Cyanothece	*Cyanothece sp.* ATCC 51142	CY51142	NC_010547
	Cyanothece sp. PCC 7425	CY7425	NC_011884
	Cyanothece sp. PCC 8802	CY8802	NC_013161
Methanosarcina	*Methanosarcina mazei strain Goe1*	MMAZ	AE008384
	Methanosarcina barkeri str. fusaro	MBAK	NC_007355
	Methanosarcina acetivorans str. C2A	MACE	AE010299
Methanococcus	*Methanococcus maripaludis strain S2*	MMARI	BX950229
	Methanococcus vannielii SB	MVAN	NC_009634
	Methanococcus aeolicus Nankai-3	MAEO	NC_009635
Frankia	*Frankia alni* ACN14a	ACN	NC_008278
	Frankia sp. CcI3	CCI3	NC_007777
	Frankia sp. EAN1pec	EAN	NC_009921
	Frankia symbiont of *Datisca glomerata*	DG	NZ_ADGT01000001*

5 strains of the actinobacteria *Frankia*. These two groups are widely apart in terms of their phylogenetic relations. Hence they are investigated separately. The members of Rhizobiales group were chosen for the comparative study to ascertain variation between the different available genomes and for better understanding of the strategies adopted by disparate strains to maintain their symbiotic apparatus. On the other hand, the *Frankia* genomes have marked differences in their size and loss-gain of several determinants (Normand *et al*., 2007). This makes them an interesting cast for comparative genomics investigation.

All nucleotide and protein FASTA files as well as GENBANK files available were utilized. All the details of the selected genomes along with their reference number are provided in the Table 3.3. Most of the further analyses were carried out in CMG-biotools which is a standalone unix platform developed for comparative microbial genomics by CBS Comparative Microbial Genomics group (Available at http://www.cbs.dtu.dk/staff/dave/CMGtools/).

3.3.2 Identification of Gene Families and Pan- and Core Genome:

To identify and process homology within and across genomes, all genes were assigned into unique gene families based on sequence similarity. The genes were translated into amino acid sequences and aligned 'all-against -all' using BLASTP (Altschul *et al*., 1990), and any two genes were considered to be a gene pair if the alignment could meet the '50/50' rule i.e. if their amino acid sequences are more than 50% identical over more than 50% of their length. Since each member of a pair can be a member of other pairs as well, all gene pairs sharing members were subsequently combined into one gene family. All genes of a genome are thus grouped into gene families. Multiple genes per genome can belong to a single gene family. A gene not finding a match with the given criteria is put in its own gene family as a singleton.

The pan and core genome plot is a simple illustration of the distribution of gene families defined above, as more and more genomes are considered. It is the result of applying a basic set theory, each genome being a set of gene family, some of which are also found in other genomes. In this context, the pan-genome becomes the union of the genomes under consideration, while the core genome is

the intersection of those genomes. The resulting pan-genome curve increases in size as more genomes are analyzed, and its shape is order-dependent, though the accumulative pan-genome is not influenced by the order of analysis. On the other hand, core genome curve decreases in size as more genomes are analyzed.

3.3.3 BLAST Matrix :

All proteomes were compared with BLASTP using the '50/50' rule to categorize genes into gene families as described above. The BLAST results were visualized in a BLAST matrix, which summarizes the results of genomic pairwise comparisons. The percentages within the matrix show the amount of proteins shared between two proteomes along with the corresponding fraction showing the number of gene families present in both genomes over the total amount of gene families in the two strains (Binnewies *et al*., 2005). For easier visual inspection, the cells in the matrix are coloured darker as the fraction of similarity increases. Hits identified within a genome are differently coloured.

3.3.4 COG Analysis:

COG is a database of proteins where each sequence is assigned to some group. All proteins within a group are believed to have a common ancestor and are likely to share a common function. The various groups are again clustered into some super-groups called functional groups (Tatusov *et al*., 2003). In this analysis, each protein was compared to the COG database using BLASTP to identify the functional groups to which they belong. An R-script was used to analyze the protein composition in core genome, dispensable set and strain specific set. The core genome, dispensable set and strain specific set together constitute the pan genome. The dispensable set includes the genes found in one or more but not all the genomes in the grouping. The strain specific set includes genes found only in particular strains.

3.3.5 Pan-genome tree vs. 16S rRNA tree:

For phylogenetic tree based on 16s ribosomal RNA, the sequences encoding 16S ribosomal RNA were predicted using RNAmmer (Lagesen *et al*., 2007) and were extracted from the set of genomes. The ClustalW (Thompson *et al*., 1994) for multiple sequence alignment was used to align the sequences. A phylogenetic tree was constructed using Bootstrap neighbour-

Table 3.3 (A) Characteristics of 15 Rhizobiales genomes used in comparative genome Study

Rhizobiales species	GBA	Code	P	GS	PCG	GC%
Azorhizobium caulinodans ORS 571	NC_009937	azo	0	53,69,772	4717	67
Bradyrhizobium japonicum USDA 110	NC_004463	bjap	0	91,05,828	8317	64
Bradyrhizobium sp. BTAi1	NC_009485	btai	1	84,93,513	7621	64
Bradyrhizobium sp. ORS 278	NC_009445	bors	0	74,56,587	6717	65
Mesorhizobium ciceri biovar biserrulae WSM1271	NC_014918	mcicer	1	66,90,028	6470	62
Mesorhizobium loti MAFF303099	NC_002678	mloti	2	75,96,297	7272	62
Mesorhizobium opportunis-tum WSM2075	NZ_ACZA 01000080	mop	0	68,54,796	6674	62
Rhizobium etli CFN 42	NC_007761	rcfn	6	65,30,228	5963	61
Rhizobium etli CIAT 652	NC_010994	rciat	3	64,48,048	6056	61
Rhizobium leguminosarum bv. Trifolii WSM1325	NC_012850	rtri	5	74,18,122	7001	61
Rhizobium leguminosarum bv. Viciae 3841	NC_008380	rvic	6	77,51,309	7143	61
Rhizobium sp. NGR234	NC_000914	rngr	2	68,91,900	6376	62
Sinorhizobium medicae WSM419	NC_009636	smed	3	68,17,576	6213	61
Sinorhizobium meliloti AK83	NC_015590	sme-lAK83	2	71,40,471	6954	62
Sinorhizobium meliloti SM11	CP001830	smelS M11	2	71,73,736	7093	62

GBA:Gen Bank Acc. No; P:Plasmid; GS:Genome size; PCG:Protein coding genes

joining method and visualized by NJ-Plot. Along with 16s RNA tree, pan-genome hierarchical clustering tree based on the gene family content has also been generated. This is a gene-content tree that illustrates similarities/ differences between genomes inside a pan-genome. The tree is constructed based on the relative Manhattan distance calculated from the BLAST matrix (Snipen and Ussery, 2010).

Two different tree i.e. “shell” and “cloud” were constructed to show different groups of genes in the pan-genome. The “shell” tree gives more weight to the genes that are present in more number than the genes present in lesser number, whereas, the “cloud” tree gives higher weight to the genes present in lesser number than the genes present in more number (Snipen *et al*., 2009; Snipen and Ussery, 2010). Perl-script present in the CMG-Biotools platform was utilized for making pan genomes trees.

3.3.6 Codon and Amino acid usage analysis:

Heatmap is a graphical display of data that uses colour codes to represent the individual values of the concerned data, thus highlighting the variations in the dataset. Heatmaps on the basis of codon usage and amino acid usage values for all the genomes under study

Table 3.3 (B) Characteristics of 5 *Frankia* genomes used in genome comparison study

***Frankia* strains**	**GBA**	**Code**	**P**	**GS**	**PCG**	**GC %**
Frankia alni ACN14a	NC_008278	acn14a	0	7,49,7934	6723	73
Frankia sp. CcI3	NC_007777	cci3	0	5433628	4548	70
Frankia sp. EAN1pec	NC_009921	ean1pec	0	8,98,2042	7191	71
Frankia sp. EuI1c	NC_014666	eul1c	0	8,81,5781	7205	72
Frankia symbiont of *Datisca glomerata*	NZ_ADGT01000001	dg	2	5,20,4281	4728	70

GBA:Gen Bank Acc. No; P:Plasmid; GS:Genome size; PCG:Protein coding genes

were plotted using CMG Biotools-package (http://www.cbs.dtu.dk/staff/dave/CMGtools/).

3.4 Homology modeling of NifH proteins:

3.4.1 Template selection and Model Building:

The amino-acid sequences of the nitrogenase iron protein (NifH) from *Bradyrhizobium* sp. ORS278 (NC_009445) and *Anabaena variabilis* (NC_007413) were retrieved from the sequence database of Integrated Microbial Genomes (img.jgi.doe.gov) (Markowitz *et al*., 2006). Thorough scanning of the sequences revealed that three dimensional structures of these proteins were not available in the Protein Data Bank (http://www.rcsb.org/pdb/home/home.do) and therefore protein 3D structure construction using homology modelling technique was undertaken.

The amino acid sequences were first subjected to PSI-BLAST (position specific iterative blast) (Altschul *et al*., 1997) for finding out the suitable template. The search was restricted to PDB structures only. Features like quality of the template structure, environmental likeness and phylogenetic similarity were considered while choosing the desired template. Alignment of the template and target sequences was carried out with the help of CLUSTAL W (Thompson *et al*., 1994). Rough three dimensional all-atom models were constructed using Modeller 9v7 (Eswar *et al*., 2003) from the sequence alignment between the template sequence and the target sequence with the parameters of energy minimization value. The technique is based upon the satisfaction of the spatial restraints acquired from the alignment (Centeno *et al*. 2005). 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 (Sen *et al*., 2010)

Secondary structure of the target protein was predicted using the web interfaces of HNN (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_hnn.html) (Guermeur, 1997) and JUFO (http://www.meilerlab.org/view.php) (Jones, 1999).

3.4.2 Model refinement:

The rough models attained by the homology modelling technique often contain certain amount of errors which may mislead if the concerned residues are present in functionally important segments (Centeno *et al*., 2005). Thus the models were further subjected to refinement. During this process constraint energy minimization was applied to all the protein atoms, using the steepest descent and conjugate gradient process to eradicate existing bad sectors between the protein atoms. All these computational analyses were performed using the Swiss-PDB Viewer package (http://www.expasy.ch/spdbv) (Kaplan and Littlejohn, 2001) *in vacuo* with the GROMOS96 43B1 parameters set. Hydrogen bonds were not considered for the final model.

3.4.3 Model evaluation:

The model was further analyzed with a series of checks for ensuring its internal stability and reliability The Auto Deposition Input Tool (ADIT) (http://deposit.pdb.org/validate) was used to construct Ramachandran Plot to look into the possible conformations of φ and ψ angles for our modeled protein. The quality of the model was checked using ProSA (Wiederstein and Sippl, 2007) (https://prosa.services.came.sbg.ac.at/prosa.php), ERRAT analysis (Colovos and Yeates, 1993) (http://nihserver.mbi.ucla.edu/ERRATv2/) and VERIFY3D (Eisenberg *et al*., 1997) (http://nihserver.mbi.ucla.edu/Verify_3D/). The refined models were then submitted to ProFunc server (http://www.ebi.ac.uk/thronton-srv/databases/ProFunc) (Laskowski *et al*., 2005) to provide functional insight of the protein and identify functionally important regions and presence of pockets in the structure were predicted using CASTp server (Dundas *et al*., 2006). Since, site-directed mutagenesis data of nitrogenase iron protein from symbiotic diazotroph was not available, *in-silico* site directed mutagenesis predictions in the functional regions were carried out using the server SDM (Worth *et al*., 2011) (http://www-cryst.bioc.cam.ac.uk/~sdm/sdm.php). I-

mutant2.0 server (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi) (Capriotti *et al*., 2005) was used to check the conformational and thermodynamic stability of the protein.

3.4.3 Conformational dynamics study:

Conformational dynamics study of the modeled protein was done with the aid of Normal Mode Analysis (NMA). NMA provides an appropriate time-independent approach to study the dynamics of relatively large proteins like nitrogenase (Tama and Sanejouand, 2001). Results from the NMAs of the protein was obtained from a number of online sources like the WEBnm@ (Hollup *et al*., 2005) (http://www.bioinfo.no/tools/normalmodes) and ElNe´mo (Suhre and Sanejouand, 2004) (http://igs-server.cnrs-mrs.fr/elnemo/index.html).These servers provides online calculation of normal modes at the residue-level.The overall dynamics of the protein can be described by a superposition of a number of linearly independent normal modes.

3.5 Molecular dynamics simulation:

3.5.1 Parameterization of metal centers:

The partial charges for the various metallic centers (Fe_4S_4 cluster) of NifH protein were calculated using quantum chemical procedures. For computational reasons, the structure centers used in these calculations were taken from the crystallographic structure without any kind of geometry optimization. The bonded parameters for the metal centers were adapted from the 43A1 GROMOS96 force field. Gaussian 03 (Frisch *et al*., 2003) was used to calculate the electrostatic potential which was fitted with RESP (Bayly *et al*., 1993) to estimate partial charges. Single-point calculations were performed using the B3LYP and 6-31G (d) basis set for all atoms. The Fe_4S_4 centers has overall +2 charges in the oxidized state and are diamagnetic.

3.5.2 Simulation procedure:

The refined homodimer model of NifH protein from *Bradyrhizobium* sp. ORS278 and monomer of NifH protein from *Anabaena variabilis* were used as starting structure for the two separate Molecular dynamics simulation. The simulations, energy minimization and trajectory analyses were carried out with the Groningen Machine for Chemical Simulation (GROMACS) program (Lindahl *et al*., 2001) running on linux platform. The starting structure was immersed in a cubic box with dimension of 20 Å for the dimer

and 10 Å in case of monomer, consisting of SPC water molecules. The adapted GROMOS96 43a1 force field parameters (Scott *et al*., 1999) were used for the simulation in both the cases. Counter ions were inserted for system neutralization The Linear Constraint algorithm (Hess *et al*., 1997) was used to constrain bond lengths and SETTLE (Miyamoto and Kollman, 1992) algorithm was used to keep water molecules rigid. The equations of motion were integrated using a time step of 2 fs. The simulation was conducted at constant temperature (300 K), coupling each component separately to a temperature bath using the Berendsen coupling method (Berendsen *et al*., 1984). Non-bonded cutoffs of 1.0 nm were used for both Lennard-Jones and Coulomb potentials.

Before running the simulation, the solvent was relaxed by energy minimization where 1000 steps of steepest descent was used followed by 1000 steps of conjugant gradient process and this was followed by 3.0 ns of simulation imposing Positional restraints (PR) on the non-H atoms at 300K. But the PR simulation run produced Linc warning at this temperature hence a new strategy was employed where the temperature for PR run was gradually increased at 50K, 100K, 200K and finally 250K before the production run were started at 300K. At each step, energy types of energy minimization i.e steepest descent, conjugant gradient and LBFGS was carried out. Finally a 3-ns simulation for homodimer was run with all protein heavy atoms position-restrained at 300 K.

The positional restraints were then subsequently released. Finally, 10 ns of final MD run for monomer and 5 ns run for homodimer were carried out and trajectories were collected. The trajectory files obtained during the whole simulation run was analyzed and RMSD, RMSF, gyration radius, potential energy were calculated with the help of GROMACS package. The time evolution of secondary structure was evaluated with the help of DSSP package (Kabsch and Sander, 1983).

3.6 *In-silico* docking of nitrogenase:

3.6.1 *Retrieval of structure:*

NifH protein model of *Bradyrhizobium* sp. ORS278 prepared earlier was chosen as the substrate macromolecule for the docking experiment. The three dimensional co-ordinates of the ADP-Ribose molecule were retrieved from NCBI-PubChem Compound database

(http://pubchem.ncbi.nlm.nih.gov/). This molecule acts as inhibitor of dinitrogenase reductase (NifH) and has role in its post translational regulation mechanism (Lowery and Ludden, 1988). Thus, ADP-ribose is considered as the ligand in the docking procedure.

3.6.2 *Molecular docking:*

The molecular docking simulation was carried out using the AutoDock 4.0 suite (Morris *et al*., 1998). In this docking simulation, semi-flexible docking protocols were followed, in which the target protein NifH was kept as rigid. The ligand being docked was kept flexible, in order to explore an arbitrary number of torsional degrees of freedom. The Autodock tool program was used to prepare, run, and analyze the docking simulation. All the rotable dihedrals in the ligand were assigned with the help of auxiliary program Auto-Tors and were allowed to rotate freely. The AutoGrid 4.0 Program, supplied with AutoDock, was used to produce a grid map for each atom type present in the ligand. This grid must surround the region of interest in the macromolecule. In the present study, the binding site was selected based on the amino acid residues which are involved in post-translational modification. Therefore, the grid was centered in the region Gly -Arg-Gly-Val-Ile-Thr. The grid box was set at 82, 60, and 50 A^0(x,y, and z). The Lamarckian genetic algorithm was chosen to search for the best conformers. The docking parameters were as follows: 50 runs, population size 150, 25,000,000 energy evaluations, 27,000 generations and RMSD-tolerance of 2.0 A^0. The docking results were analyzed in Autodock tools. A perl script was used to separate various substrate-ligand docked conformations.

3.7 Comparative phylogeny:

3.7.1 Sequence based phylogeney:

3.7.1.1 Sequence Retrieval:

Amino acid sequences of NifH, NifK and NifD proteins from different diazotrophs and Cyanobacteria were retrieved from from Integrated Microbial genomes database (www.img.jgi.doe.gov) (Markowitz *et al*., 2006). The organisms included in the analysis are listed in Table 3.4 . A Total of 50 amino acid sequences were utilized in the phylogenetic analysis.

3.7.1.1 Mutliple alignment and Tree generation:

The ClustalW program was used multiple amino acid sequences alignment (Thompson *et al*., 1994) using default gap penalties.

Table 3.4 List of organisms whose Nif proteins are used for Phylogenetic analysis based on sequence based and Condensed matrix based method. (*denotes draft sequences)

Organism Name	Accession No.
Cyanobacteria	
Nostoc sp. PCC 7120 †	NC_003272
Nostoc azollae	NC_014248
Cyanobacterium UCYN-A	NC_013771
Anabaena variabilis ATCC 29413	NC_007413
Trichodesmium erythraeum IMS101 †	NC_008312
Lyngbya sp. PCC 8106 †	NZ_AAVU01000084*
Cyanothece sp. PCC 7424 †	NC_011738
Cyanothece sp. PCC 7425 †	NC_011884
Synechococcus sp. JA-3-3Ab †	NC_007775
Proteobacteria	
Azoarcus sp. BH72 †	NC_008702
Azotobacter vinelandii DJ †	NC_012560
Beijerinckia indica †	NC_010581
Klebsiella pneumoniae 342 †	NC_011283
Burkholderia xenovorans	NC_007951
Burkholderia vietnamiensis	NC_009256.
Azospirillum sp. B510 †	NC_013854
Bradyrhizobium japonicum USDA 110 †	NC_004463
Bradyrhizobium sp. ORS278 †	NC_009445
Azorhizobium caulinodans ORS 571 †	NC_009937
Mesorhizobium loti MAFF303099 †	NC_002678
Rhizobium leguminosarum bv. viciae 3841 †	NC_008378
Rhizobium leguminosarum bv. trifolii WSM1325 †	NC_011366
Sinorhizobium medicae WSM419	NC_009620
Sinorhizobium meliloti 1021	NC_003037
Rhodopseudomonas palustris BisB18 †	NC_007925
Rhodobacter sphaeroides ATCC 17029 †	NC_007493
Geobacter sulfurreducens PCA †	NC_002939
Acidithiobacillus ferrooxidans ATCC 53993 †	NC_011206
Desulfovibrio vulgaris subsp. vulgaris DP4	NC_008751
Halorhodospira halophila SL1	NC_008789.
Methylococcus capsulatus str. Bath †	NC_002977
Desulfuromonas acetoxidans DSM 684 †	NZ_AAEW00000000*
Methylobacterium nodulans ORS 2060	NC_011894
Desulfatibacillum alkenivorans AK-01	NC_011768
Actinobacteria	
Frankia sp. CcI3 †	NC_007777
Frankia sp. EAN1pec †	NC_009921
Frankia alni ACN14a †	NC_008278
Frankia DG †	NZ_ADGT01000001*
Frankia sp. EUN1f †	NZ_ADGX01000001*
Firmicutes	
Clostridium pasteurianum †	taken from PDB id 1CP2
Clostridium acetobutylicum ATCC 824 †	NC_003030
Clostridium botulinum ATCC 3502 †	NC_009495
Desulfitobacterium hafniense Y51	NC_007907
Alkaliphilus metalliredigens QYMF †	NC_009633.
Green sulphur bacteria	
Chlorobium ferrooxidans DSM 13031 †	NZ_AASE00000000*
Chlorobium limicola DSM 245 †	NC_010803
Chlorobium tepidum TLS †	NC_002932
Aquificae	
Hydrogenobacter thermophilus TK-6 †	NC_013799
Methanogenic Archaea	
Methanothermococcus okinawensis	NC_015636.1
Methanococcus vannielii SB †	NC_009634

Phylogenetic analyses were performed using the neighbour-joining (NJ) method (Saitou and Nei, 1987) implemented in MEGA4 program (Tamura *et al*., 2007) with 1,000 bootstrap replicates. .

3.7.2 *Nucleotide triplet based phylogeny:*

3.7.2.1 Retrieval of gene sequences:

The nucleotide sequences of *nif*H, *nif*K and *nif*D genes from different N2 fixing microorganism including cyanobacteria, proteobacteria, actinobacteria and archea (Listed in the Table 3.4) were retrieved from Integrated Microbial genomes database (www.img.jgi.doe.gov) (Markowitz *et al.,* 2006).

3.7.2.2 Frequency of triplets of nucleic acid bases:

The DNA sequences comprises of four bases i.e. A,T,G and C which can form 64 possible triplets. The triplets contain all the relevant information for polypeptide biosynthesis. A program written in Trubo C++ written in our lab was utilized to count each and every potential triplet from nucleotide sequences of *nif*H, *nif*K and *nif*D from the studied microbes. 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 (Sur *et al*., 2010) each of which containing all entries of the cubic matrix (Randic *et al*., 2001). Usually the group of 4 × 4 matrices {M1, M2, M3, M4} is the representative of the cubic matrix. The four matrices contain not only the information about frequencies of occurrence of all triplets of a DNA sequence but also the information about the frequencies of occurrence of pairs and every letter in a DNA sequence. 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 *et al*., 2001). In our method the codon positions 1, 2 and 3 were given equal weight subsequently addition or deletion of bases that have had occurred during the course of evolution were given due care so as to incorporate their influence.

3.7.2.3 Calculation of eigen value and

construction of Phylogram:

Leading eigenvalues of the matrices were calculated using MATLAB (version 7.10.0) software. These eigenvalues are a special set of scalars associated with a linear system of equations (i.e., a matrix equation) that are sometimes also known as characteristic roots, characteristic values (Hoffman and Kunze, 1971), proper values or latent roots (Marcus and Minc, 1988). Evaluation of DNA sequences for similarity or dissimilarity is normally aided by the convenience of leading eigenvalues calculated by this method. Distance matrices of the studied sequences were constructed by summing up the square of the difference of eigen values. Phylogram were built by cluster analysis of the similarity matrix using PHYLIP (ver 3.65) (Felsenstein, 1989) and drawn with PHYLODRAW (Ver 0.8) (Choi *et al*., 2000).

3.7.3 Structure based phylogeny:

3.7.3.1 Search for the related 3D protein structure:

Structural coordinates for members of the extended NifH protein family were extracted from the Protein Data Bank (PDB) (Bernstein *et al*., 1977) using pairwise 3D alignments calculated by the DALI program (Holm and Sander, 1995). A detailed list of the various protein 3D structures considered for phylogenetic analysis is provided in

Table 3.5 Details of the NifH/Bchl Family 3D protein structures used for phylogenetic analysis.

ID	Details about the protein	Strain
1NIP (PDB id)	Nitrogenase iron protein	*Azotobacter vinelandii*
1G5P (PDB id)	Nitrogenase iron protein (All-ferrous [4Fe-4S]0 form)	*Azotobacter vinelandii*
1CP2 (PDB id)	Nitrogenase iron protein	*Clostridium pasteurianum*
1G1M (PDB id)	Nitrogenase iron protein (All-ferrous [4Fe-4S]0 form)	*Azotobacter vinelandii*
2AFH (PDB id)	Nitrogenase complex	*Azotobacter vinelandii*
3NFP; 3NHP; 3NGP	Nitrogenase iron protein #(Sen *et al*. 2010)	*Frankia* ACN14a, *Frankia* CcI3 and *Frankia* Ean1pec
TRI	Nitrogenase iron protein # (Zehr *et al*. 1997)	*Trichodesmium* spp.
BRA	Nitrogenase iron protein from # (Thakur *et al*. 2012)	*Bradyrhizobium* sp. ORS278
1ANA	Nitrogenase iron protein ‡	*Anabaena variabilis*
3FWY (PDB id)	L -protein of light-independent protochlorophyllide reductase (BchL)	*Rhodobacter sphaeroides*
1HYQ (PDB id)	Crystal structure of bacterial cell division regulator MinD.	*Archaeoglobus fulgidus*
2BEK (PDB id)	Bacterial chromosome segrega-tion protein Soj	*Thermus thermophilus*

('#' stands for validly published protein Homology Model; '‡' stands for Homology model id Q3M4L0 taken from SWISS-MODEL repository)

Table 3.5. Since x-ray crystallographic and NMR structures of many symbiotic diazotrophs and cyanobacteria is still missing in PDB, therefore we have included some of the validly published protein structures deduced by computational tools. The DALI algorithm was also used for the pairwise comparison and alignment of three dimensional structures of NifH protein and its relative neighbors. The DALI- root mean square distance (the square root of the average squared Euclidean distance over all topologically equivalent pairs of alpha-carbon positions) was used as indicators of protein similarity, resulting in distance matrices. The matrix was used to create phylogenetic (similarity) tree using the Fitch-Margoliash algorithm (Fitch and Margoliash, 1967) as implemented in the FITCH program of the PHYLIP package (Felsenstein 1989).

3.8 Functional divergence in NifH/ BChl protein family:

3.8.1 Data collection and multiple sequence alignments:

Amino acid and nucleotide sequences of NifH protein from 40 different diazotroph (marked as "Ŧ" in Table 3.4) were obtained from JGI-IMG database (www.img.jgi.doe.gov). This comprises representatives bearing the three major types of NifH enzyme. Protein BLAST was carried out to identify potential homologues of the NifH from light independent protochlorophyllide (Pchlide) reductase. Protein was considered as homologue when amino acid identity was above 35% over a stretch of ≥ 150 amino acids. Five such sequences were selected for further analysis. This includes bchL protochlorophyllide reductase sequences from bacteria like *Rhodopseudomonas palustris* CGA009*, Erythrobacter sp.* NAP1*, Roseobacter denitrificans OCh 114, Rhodobacter sphaeroides, Rhodobacter capsulatus* SB1003. Total of 45 protein sequences were subjected to multiple alignment using ClustalW 1.83 (Thompson *et al*., 1994). A phylogenetic tree was constructed based on neighbor joining method using the software MEGA 4 (Tamura *et al*., 2007).

3.8.2 Functional divergence detection:

NifH sequence duplication events were tested for type I functional divergence based on the method by Gu (1999; 2001). The analysis was carried out with Diverge version 2.0 program. This method is based on maximum likelihood procedures to estimate

significant changes in the rate of evolution after the emergence of two paralogous sequences. Type I sites represent amino acid residues conserved in one subfamily but highly variable in another, implying that these residues have been subjected to different functional constraints. Previously created NJ tree was utilized in this analysis. The coefficient of functional divergence (θ) and the posterior probability for the functional divergence were calculated for each position in the alignment. To detect amino acid residues reflecting functional divergence, NifH subfamilies were pair-wise compared to each other. The cut-off value for the posterior probability was determined by consecutively eliminating the highest scoring residues from the alignment until the coefficient of Functional divergence dropped to zero. Residues predicted to be functionally divergent were finally mapped onto the three dimensional model of nitrogenase iron protein.

3.8.3 Detection of Positive selection:

DNA sequences along with the multiple alignments of their corresponding protein sequence were uploaded to the PAL2NAL web server (Suyama *et al*., 2006) which converts them into a codon alignment. Phylogenetic tree reconstruction was done using maximum likelihood method (ML) from the protein alignment using PhyML software package (Guindon and Gascuel, 2003) with the gamma distribution model implemented to account for heterogeneity among sites. The best fit model of protein evolution used for this purpose was the JTT+I+G model. Subsequently, the codon alignment and ML tree were provided to CODEML (Yang, 1997) program under PAML package to uncover the possibility of positive selection based on site-specific divergence models. The commonly used commonly used PAML models are M1a (neutral model) and M2a (positive), and M7 (neutral) and M8 (positive). These models allow the ratio of nonsynonymous to synonymous substitutions (dN/dS or ω) to vary across sites; ω of each site in the gene is considered a random variable drawn from a statistical distribution. For the detection of positive selection in NifH/ BchL family, site-specific likelihood models M7 and M8 were applied to the sequences with the appropriate tree topology (Nielsen and Yang, 1998; Yang *et al*., 2000). Model M7, which does not allow for sites under positive

selection, was compared to model M8, which allows for sites under positive selection. M7 assumes a beta distribution for ω between 0 and 1 over all sites, while M8 adds an additional site class (ω ≥ 1), with ω estimated from the data. A likelihood ratio test (LRT) was performed by calculating two times the log likelihood values and comparing this value to a $\chi 2$ distribution with two degrees of freedom. Posterior probabilities of codons under positive selection were computed in model M8 using Bayes empirical Bayes when the LRT was significant.

3.9 Statistical coupling analysis:

3.9.1 Sequence retrieval and Multiple Sequence Alignment:

Forty five protein sequences comprising of members of NifH/Bchl protein family were considered for the SCA analysis. The sequences were same as the ones utilized in the functional diveregence analysis (section 3.8). The sequences were then subjected to multiple sequence alignment (MSA) using ClustalW (Thompson *et al*., 1994). Alignment positions with less than 20% gaps we considered for SCA calculation.

3.9.2 Statistical coupling analysis:

The analysis of conservation and pairwise correlation in the multiple sequence alignment uses updated versions of the SCA method (Lockless and Ranganathan, 1999; Suel *et al*., 2003). Lockless and Ranganathan, (1999) introduced a sequence-based method to investigate statistical interactions between residues (Statistical Coupling Analysis (SCA). Later Halabi *et al*., (2009) grouped these statistically correlated amino-acids into quasi-independent groups called sectors and studied their characteristics in Serine proteases

The SCA code was obtained from Rama Ranganathan's Lab through personal communication. This includes a MATLAB script for reproduction of all of the calculations which can be implemented in MATLAB (Mathworks) toolbox. Default SCA parameters are used and the pairwise positional correlation varies from 0.2 – 0.6 majorly. All the SCA results are mapped back to the 3D protein structure NifH from *Azotobacter vinelandii* (PDB id 1G5P) for ease of structural mapping and comparison with structure based analyses. Thus, protein sector was indentified and their respective positions were mapped in the 3D protein structure.

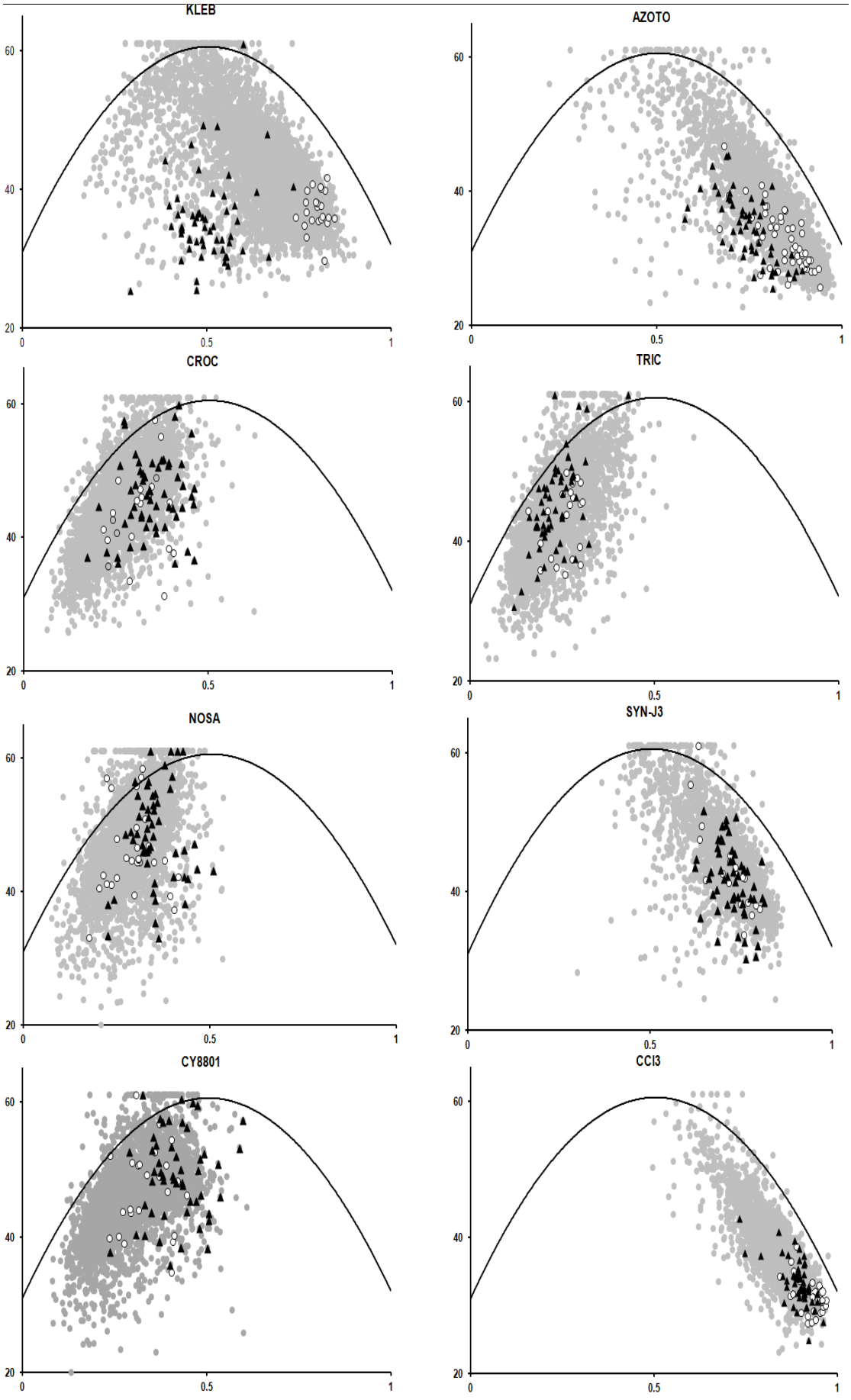
Chapter 4
Results and Discussion

4.1 Codon and Amino acid usage analysis:

4.1.1 Synonymous codon usage pattern and heterogeneity:

The main motive of this study on the codon usage patterns was to estimate the level of heterogeneity in codon use. The genomes of various nitrogen-fixing microorganisms were analyzed for codon heterogeneity using the parameters like GC content, GC3 content, Nc, CBI and Fop. The effective number of codons (Nc) versus GC3 plots has been suggested to be an important means of investigating codon usage variations among genes in the same genome (Peden, 1999). The Figure 4.1 depicts the Nc/GC3 plots for various diazotrophic genomes. In actinobacterial *Frankia* strains, the effective number of codon values range from 21 ± 2 to 61 ± 0 for all five genomes suggesting that these highly GC-rich genomes exhibited considerable heterogeneity in codon usage. The location of different categories of genes for *Frankia* CcI3 has been shown in the Nc/GC3 plot (Figure 4.1). Ribosomal protein genes are strongly clustered at lower end of the plots indicating a significantly strong codon bias in these genes. The genes associated with nitrogen fixation, are also shown in the plots. Genes associated with nitrogen fixation are more or less clustered strongly with the ribosomal protein genes. Other *Frankia* strains under consideration also have similar pattern of distribution of genes in Nc/GC3 plot (See the Supplementary Figure SF1.1).

In *Chlorobium*, most of the ribosomal protein genes were loosely clustered in the plot and remain below the expected curve (Figure 4.1). This indicates lower codon bias in these genes. The nitrogen -fixing genes remain clustered more or less strongly in all the *Chlorobium* genomes. In cases where synonymous codon bias is completely dictated by GC3s, Nc values fall on the expected curve in the Nc vs GC3 plot (Sur *et al.*,



Continued to the next page

Continued from previous page

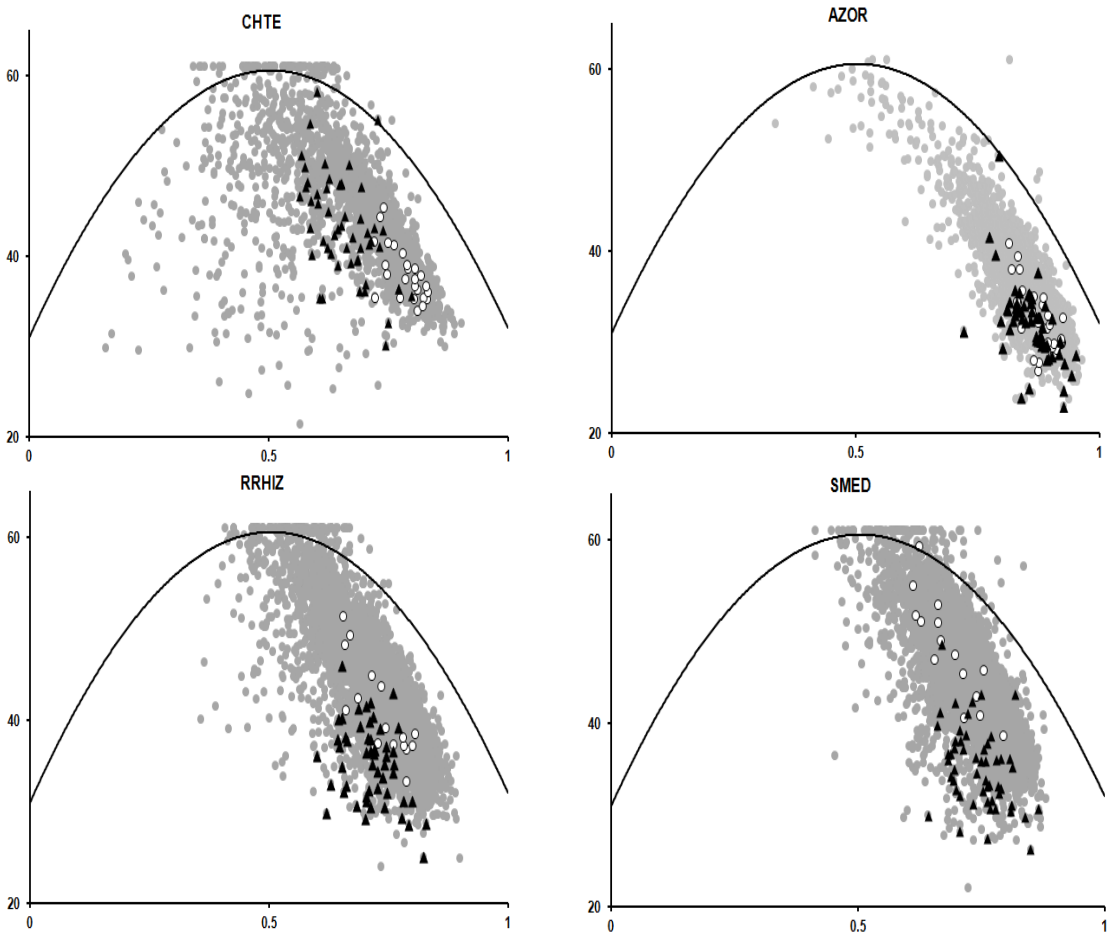


Figure 4.1 Nc/GC3 plots for KLEB, AZOTO, CROC, TRIC, NOSA, SYN-J3, CY8801, CCI3, CHTE, AZOR, RRHIZ and SMED(Left to right and top to bottom). Protein coding genes are represented by grey circle, ribosomal protein genes by black circle and nitrogen fixing genes by white circles. The continuous curve in each plot symbolizes the null hypothesis that the GC bias at the synonymous site is solely due to mutation but not selection. X and Y axis are represented by GC3 and Nc respectively.

2008). However, in the *Chlorobium* genomes it is seen that barring a few, majority of the genes were below the expected curve. This is a reflection of the fact that synonymous codon bias in majority of the genes in *Chlorobium* is quite independent of the overall base compositions.

In the cyanobacterial strains, the GC, GC3 and Nc values varied in different species. Many of the studied cyanobacterial strains were moderate to low GC containing and have high Nc values indicating low codon bias. In Nc/GC3 plot barring *Synechococcus* and *Trichodesmium*, ribosomal protein genes for most of the other strains remained scattered indicating that they had dissimilar codon bias while in the former two, they were clustered at one end of the plot. The NFGs for the cyanobacterial strains also remained

scattered with the exception of *Synechococcus* and *Trichodesmium*. The Nc/GC3 plots for some representative proteobacterial members are depicted in Supplementary Figures SF1.3-1.4. In *Azorhizobium, Bradyrhizobium* ORS278 the RPGs along with the NFGs were clustered at clustered at the lower end of the plot and they are therefore highly biased. But in other members of Rhizobiales like *Rhizobium leguminosarum* bv. *trifolii* WSM2304, *Rhizobium rhizogenes* K84, *Sinorhizobium medicae* etc, the RPGs remain moderately clustered at lower spectrum but the NFGs remain scattered.

The mean values of different indices used for the study of codon usage patterns in diazotroph genomes are provided in Supplementary Table ST1. The values are provided for three set of genes i.e. Protein coding genes (PCG), Ribosomal protein genes (RPG) and nitrogen-fixation related genes (NFG). As expected, a correlation between effective number of codons and GC3 was observed: effective number of codons values decreased with a corresponding increase in GC3 values and vice-versa. For example, in actinobacterial diazotroph *Frankia*, the GC, GC3 content was very high but their Nc values were low. These low effective number of codons values indicated a high degree of codon bias. Ribosomal protein genes and genes associated with nitrogen fixations had a lower mean effective number of codons value than the mean value obtained for all of the protein encoding genes for all of the genomes. Amongst the *Frankia* strains under consideration, there is not much difference in the CBI and Fop values. However, moderate Fop values indicated that other than mutational pressure additional factors may play a role in effecting codon usage variations in these genomes. In *Cholorobium*, the GC and GC3 content is moderate but the Nc values were quite high symbolizing low codon usage bias.CBI values were very low representing low codon bias. The Fop values are moderate however RPGs and NFGs had higher Fop values indicating higher percentage of optimal codons in these genes. Cyanobacterial strains have moderate to low GC content but most of them have high Nc values indicating low codon bias. Low CBI values and in some cases negative CBI values indicates lower levels of optimal codons. The aforesaid fact is further substantiated by the moderate Fop

values.

4.1.2 Expression pattern analysis:

The mean values of CAI across the genomes of several diazotroph are provided in the Supplementary Table ST1 along with other indices. The CAI values for *Frankia* genomes are quite high with the RPGs having higher mean values CAI values compared to protein coding genes. Nitrogen fixation related genes also had high CAI values however they were comparatively lower than protein coding genes. The *Frankia* EAN1pec had the highest average CAI values (around 0.83) for NFG's amid the *Frankia* strains. The CAI values for cyanobacterial strains revealed high expression levels and varied at the interspecific level. In most cases ribosomal protein genes had higher CAI values as expected. The CAI values of NFGs were high in most of the studied strains. *Trichodesmium* had the highest average CAI value for NFGs amongst the cyanobacteria, which is at par with that of its ribosomal protein genes while the *Synechococcus* strains had the lowest average CAI value for NFGs. The CAI values for the proteobacterial members are also quite high. As expected RPGs are more highly expressed compared to other studied categories of genes but in majority of cases the difference is little. The overall CAI values of the genomes were correlated with Nc, GC and GC3 values and correlation coefficient is provided in the Table 4.1. CAI values did not reveal any significant trend in correlation with Nc and GC content. In fact, in many of the studied genomes like *Cyanobacterium* UCYN-A, *Anabeana variabilis*, *Azotobacter vinelandii* and even in *Frankia* strains a negative correlation is noticed between Nc and CAI values. CAI also showed positive correlation with GC values in most of the symbiotic proteobacterial strains like *Bradyrhizobium, Synorhizobium, Rhizobium* etc. In most of the cyanobacterial strains, CAI values showed negative correlation with GC value. However, in *Synechococcus* strains with GC content close to 60%, it was noticed that both GC and GC% is positively correlated with CAI. Barring a few cases of AT-rich Cyanobacterial genomes like *Anabeana variabilis, Cyanothece strains, Lygbya* most of the studied strains revealed moderate positive correlation of CAI with GC3 content (at ($P<0.001$) level implying that highly expressed genes have higher GC3 content.

The CAI values for the protein coding

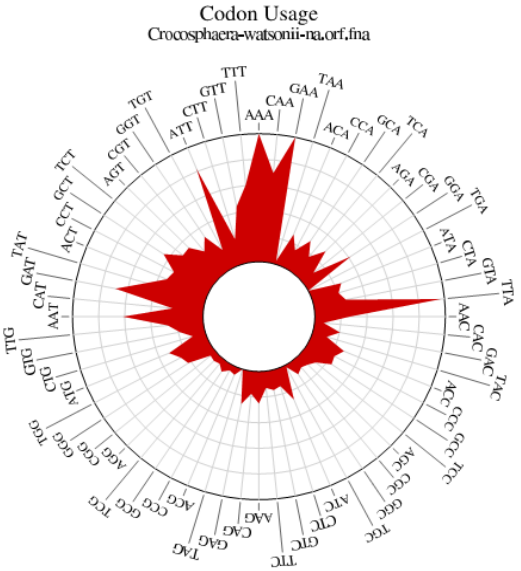
Table 4.1 Correlation of CAI with Nc, GC and GC3 values in various diazotroph genome

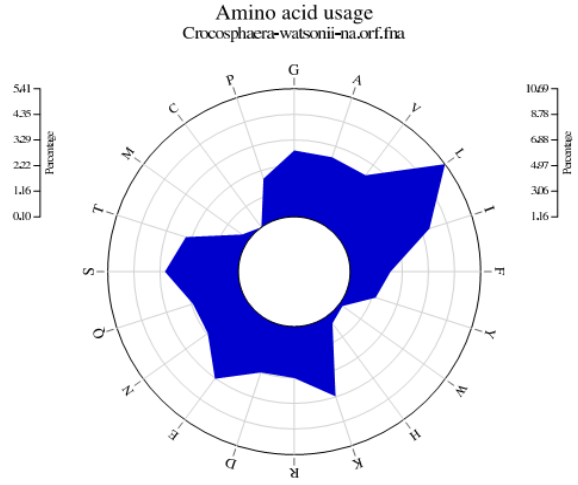
Organism	Nc/CAI	GC/CAI	GC3/CAI
CYAN	-0.54	-0.28	-0.77
CROC	-0.54	-0.32	-0.51
SYN-J3	-0.73	0.57	0.85
SYN-J2	-0.65	0.50	-0.79
NOSP	-0.45	-0.48	-0.63
NOSA	-0.37	-0.27	-0.52
NOS7120	-0.25	-0.29	-0.45
NOD	0.02	0.03	0.02
ANA	-0.38	-0.28	-0.39
LYN	-0.29	-0.44	-0.21
CY7822	-0.42	-0.23	-0.46
CY51142	-0.43	-0.25	-0.47
CY7424	-0.50	-0.29	-0.53
CY8801	-0.25	0.031	-0.037
TRIC	-0.65	-0.64	-0.71
RVICIA	-0.46	0.19	0.58
RTRI	-0.85	0.49	0.80
RCIAT	-0.54	0.34	0.70
RCFN	-0.27	0.20	0.29
RNGR	-0.74	0.52	0.65
RRHIZ	-0.67	0.55	0.54
BJAP	-0.19	0.13	0.20
BTAI	-0.20	0.14	0.22
BORS	-0.10	0.02	0.52
MCICER	-0.35	0.36	0.42
MLOTI	-0.39	0.39	0.40
MOP	-0.46	0.27	0.66
SMED	-0.30	0.18	0.25
SMEL	-0.45	0.15	0.53
AZOR	-0.22	0.15	0.31
AZOTO	-0.83	0.51	0.82
KLEB	-0.14	0.08	0.10
BEJ	0.14	-0.33	-0.19
ACN	-0.19	0.18	0.23
CCI3	-0.53	0.32	0.61
EAN	-0.53	0.38	0.58
DG	-0.82	0.50	0.94
EUN1f	-0.68	0.45	0.57
CHLO	-0.29	0.23	0.16
CHTE	-0.04	0.19	0.25

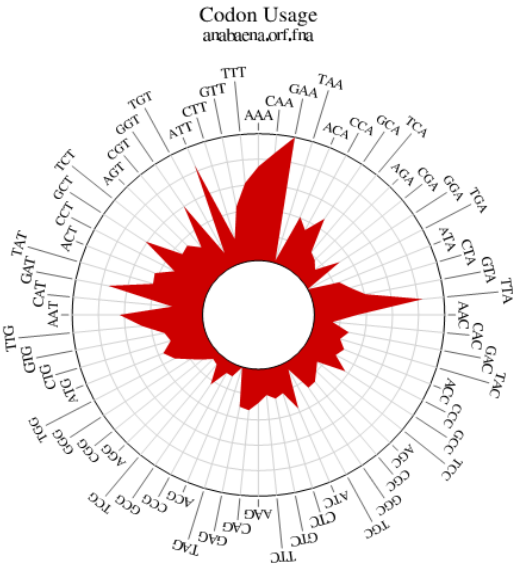
genes of *Frankia* ACN14a was correlated with microarray expression data (in terms of fluorescence values) acquired for symbiotic and pure cultures in two media (FrBAP+, FrBAP-) and in symbiotic tissues of *Alnus glutinosa.* This work was done to see whether the calculated CAI value has any indeed any relation with the experimentally acquired expression data. The results obtained in this study showed that there is no apparent trend in the relation of CAI and fluorescence values obtained in the two media. In case of the fluorescence values obtained in the symbiotic tissues of *Alnus glutinosa* there is slight hint of positive correlation (with Pearson co-relation coefficient R = 0.0125). However before concluding anything from this it must be kept in mind that microarray expression data varies with situation like age, time of collection, place and even the medium of culture because expression of protein varies with situation. So it is difficult as well as unreasonable to find a correlation with calculated CAI values with microarray data. CAI can only be a guiding tool in estimating the potential expression pattern based on reference set. CAI has been shown previously to have a correlation with protein expression levels determined by mass spectrometry in certain organism (Wang *et al.*, 2005).

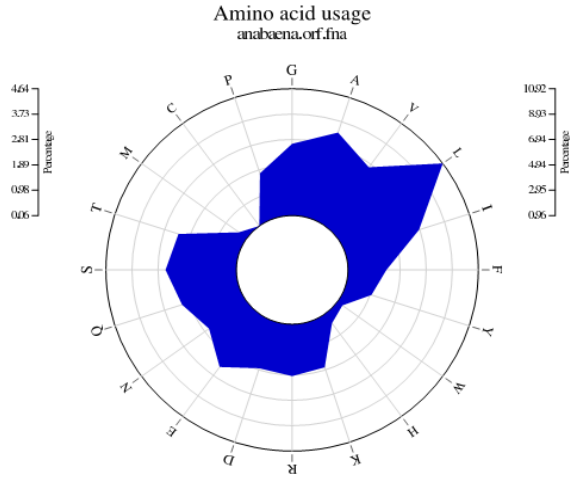
4.1.3 Rose-plots of Codon usage and Amino acid usage:

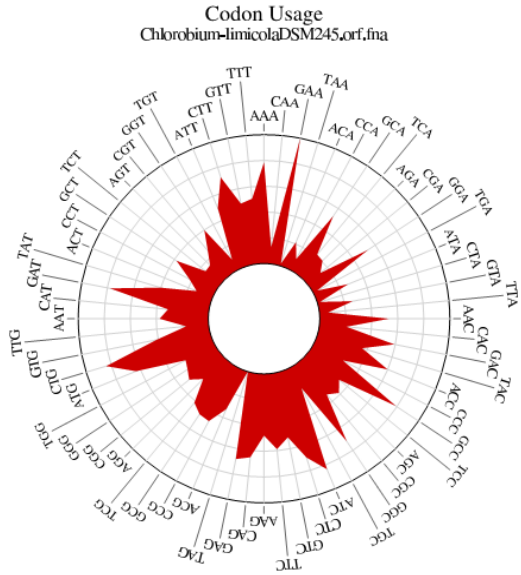
The rose plot provides a graphical outline of differences in overall codon usage and amino acid usage for various nitrogen-fixing microbes. Codon/ Amino acid usage rose plots of some of the representative diazotrophs from different lineages are provided in Figure 4.2 and rest of the plots are provided in the Supplementary Figures SF2.1-2.10. In the Rose-plot of codon usage, all of the 64 codons are listed in the perimeter and the frequency of each codon is drawn on a radial scale. The frequency of each codon is plotted in red. In the rose plot of amino acid usage, frequency of each amino acid is plotted in blue. Distinguishable signatures for each genus are apparent. AT-rich genomes of *Cyanobacterium* UCYN-A, *Crocosphaera watsonii and Trichodesmium erythraeum, Cyanothece strains* show a bias at the third codon position with a strong preference towards bases A and T in the third position. Frequency of codons like TTA, AAA, ATT etc are quite higher in these genomes. Reverse trend can be found in the GC rich genomes

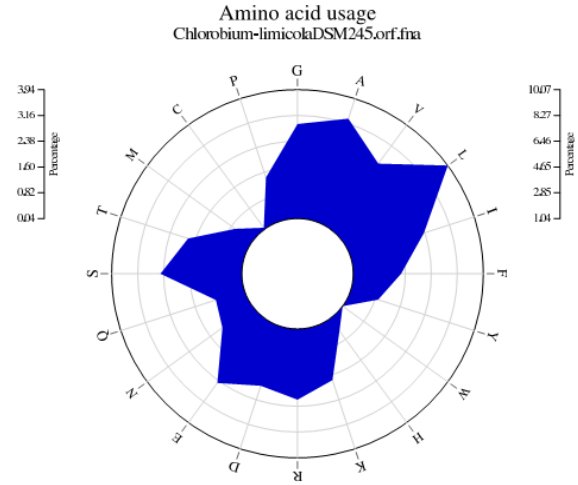






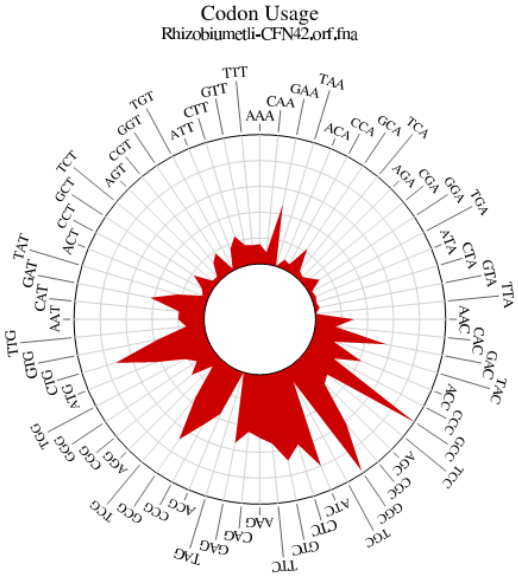


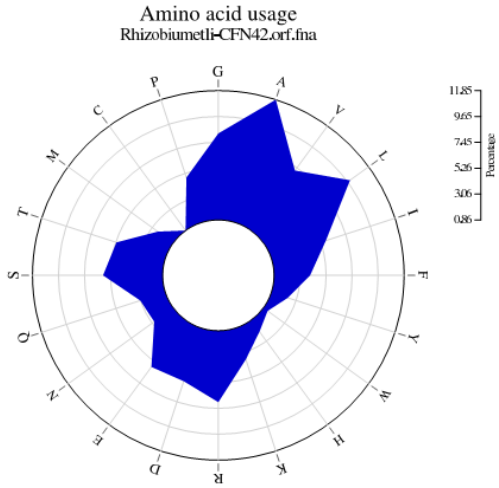


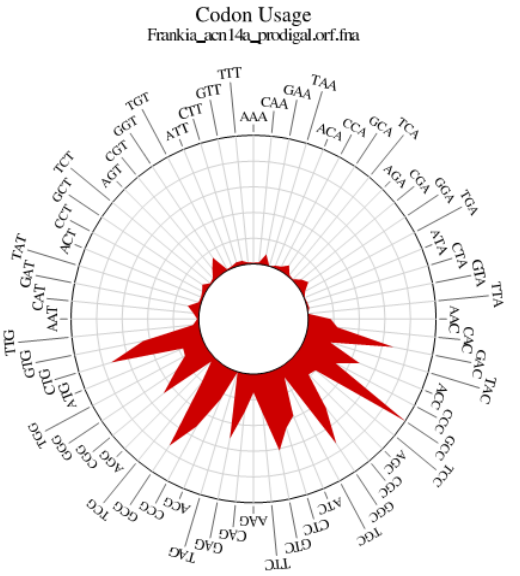


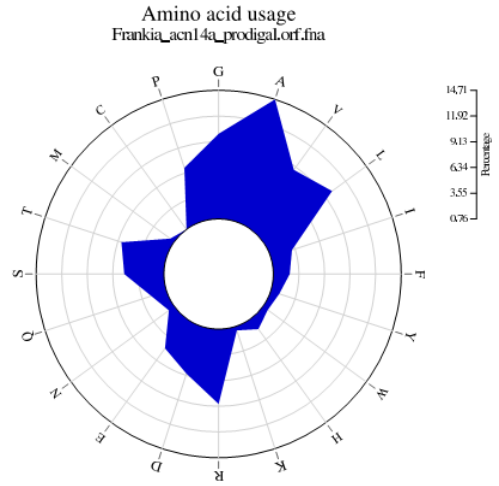
Continued to next page

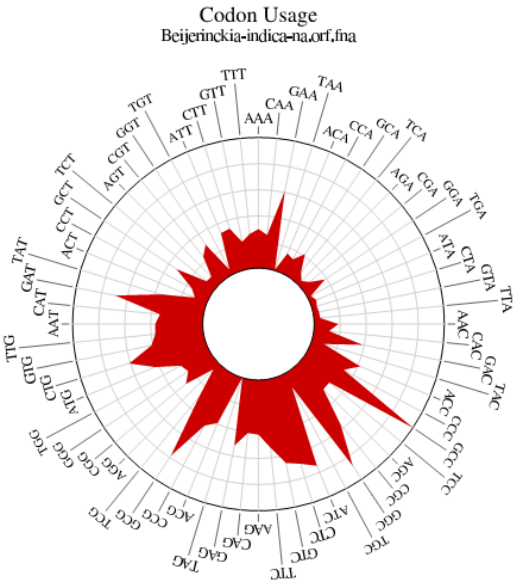
Continued from previous page

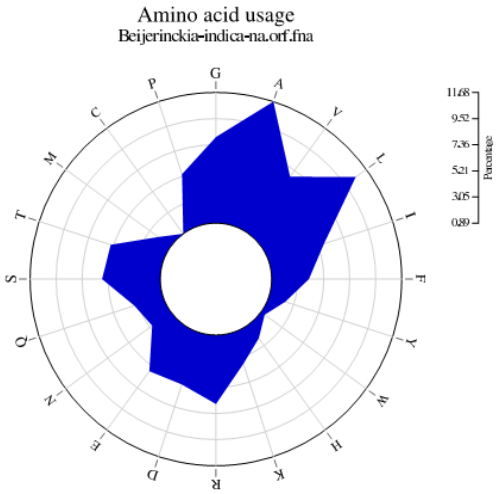












Continued to next page

Continued from previous page

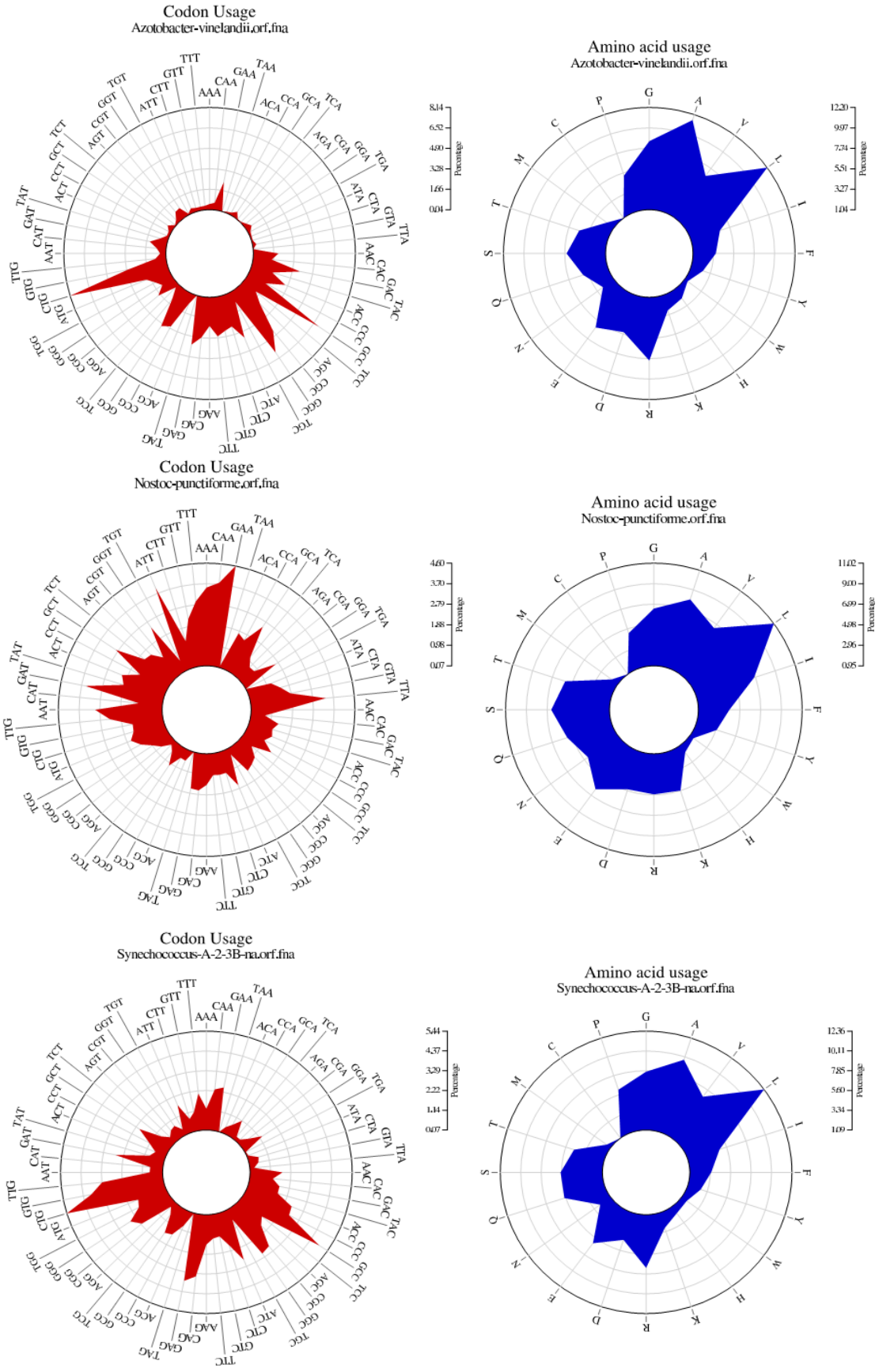


Figure 4.2 Codon usage and Amino acid rose plots for representatives of different diazotrophs. The frequency scale is represented at the right side of the rose plots

of *Frankia* and *Azotobacter vinelandii,* which frequently use codons that tend to end in either C or G. In such cases, codons like GCC, GCG, GGC etc are the most preferred ones. The Rhizobiales members under consideration have GC% content ranging from 61-67%, hence they also show a subtle bias for codons ending with C or G. A different trend is noticed in genomes of *Chlorobium* strains*, Beijerinckia indica* and *Klebsiella pneumonia.* These genomes do not have a preferential codon usage and does not show a bias at the third codon position. In fact, the three most commonly used codons of *Chlorobium limicola* genome are GAA, CTG, ATC respectively. The amino acid usage pattern also reveals some amount of variation across the genomes. In most cases, leucine is the most commonly used amino acid. Exception is seen in the genome of *Frankia* strains and members of Rhizobiales, where alanine is the most favored amino acid. In genomes of *Beijerinckia indica and Azotobacter vinelandii,* it was found alanine and leucine is almost equally utilized. In spite of the variation in amino usage pattern, nothing can be substantially concluded from the amino acid usage plot because an AT-rich genome and a GC-rich genome might code for a similar amino acid composition, even though it has different third position bias. Thus codon usage rose plot are better in providing a general knowhow of the pattern and behavior of codon usage from a simple pictorial representation.

4.1.4 Correspondence analysis of amino acid usage:

Correspondence analysis was applied to the amino-acid composition matrix (see Supplementary Table ST2). The resulting distribution of amino acids, along with the surveyed species, the illustrative variables (char, pol, hyd) is shown in Figure 4.3. This Figure shows the organisms distribution on the factorial space defined by the first (F1) and the second (F2) factors. As indicated on the Figure, F1 and F2 represent 84.4% and 8.84% respectively of the total information included in the analysed data table. According to the amino-acid composition, 3 main clusters are noticed. One of the clusters comprises of symbiotic proteobacteria like *Rhizobium, Sinorhizobium, Mesorhizobium* etc. The free-living N2 -fixer *Beijerinckia indica* also belong to this cluster. The other cluster exclusively comprises of strains of

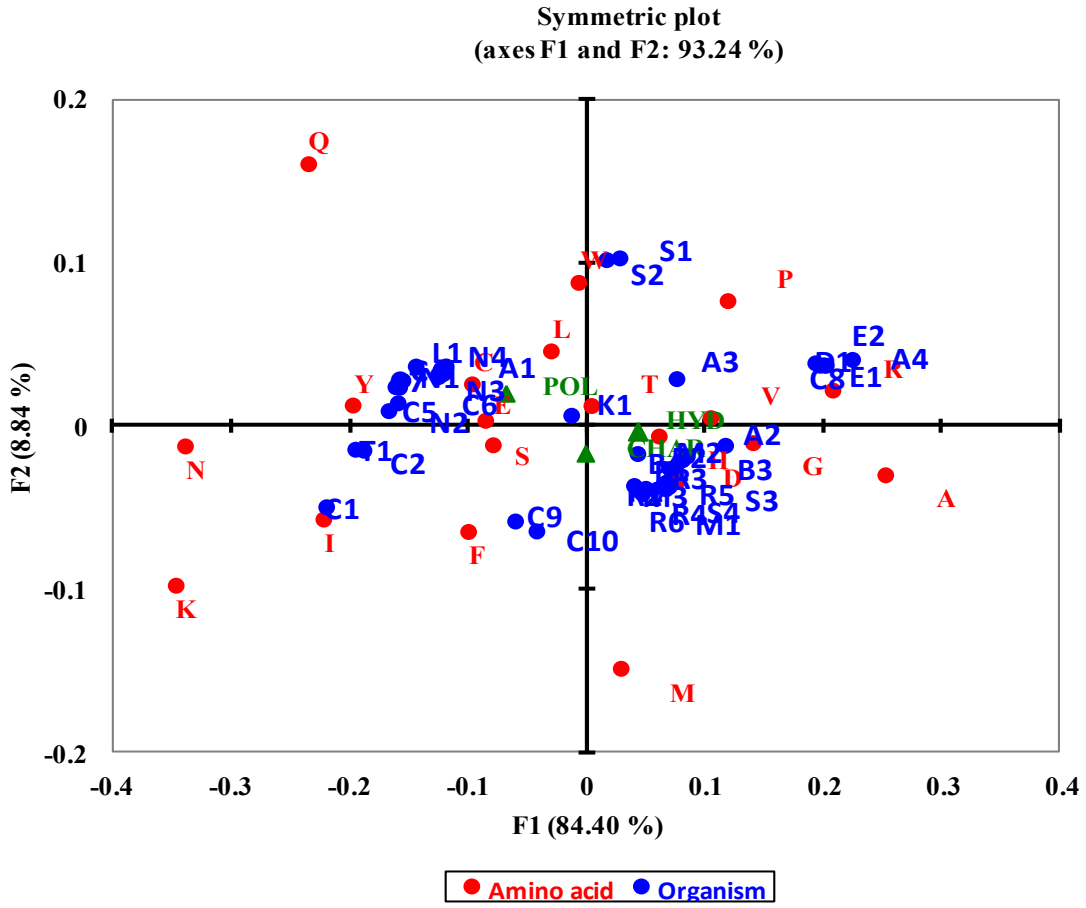


Figure 4.3 Distribution of the surveyed species according to their relative amino-acid composition, as obtained on the first factorial space (F1 and F2) by correspondence analysis. (C1=CYAN, C2=CROC, S1=SYN-J3, S2=SYN-J2, N1=NOSP, N2=NOSA, N3=NOS7120, N4=NOD, A1= ANA, L1=LYN, C4=CY7822, C5=CY51142, C6=CY7424, C7=8801, T1=TRIC, R1=RVICIA, R2=RTRI, R3=RCIAT, R4=RCFN, R5=RNGR, R6=RRHIZ, B1= BJAP, B2=BTAI, B3=BORS, M1=MCICER, M2=MLOTI, M3=MOP, S3=SMED, S4=SMEL, A2=AZOR, A3=AZOTO, K1= KELB, B4=BEJ, A4= ACN, C8=CCI3, E1= EAN, D1=DG, E2=EUN1f, C9=CHLO and C10=CHTE, respectively)

nitrogen fixing actinobacteria *Frankia*. Besides, another noticeable cluster comprising of various Cyanobacteria like *Cyanothece, Nostoc, Anabaena* etc. were also found. However, marine cyanobacteria like *Cyanobacterium* UCYN-A, *Crocosphaera watsonii* and *Trichodesmium erythraeum* doesn't belong to this cluster and are present in a different quadrate along with thermophilic green-sulfur bacteria *Chlorobium*. Free-living nitrogen fixers like *Azotobacter, Klebsiella* don't belong to any distinct cluster. One apparent trend in this figure is that symbiotic proteobacteria are situated near to the charged amino-acids (char), which suggests that these organisms encode on average, higher levels of charged amino-acids. These organisms are mainly defined by the relatively high values of Asparatic acid and

Histidine. On the other hand, Cyanobacteria such as *Cyanothece, Lynbya* etc encode higher level of polar residues (pol). This is mainly due to the relatively high values of Cysteine, Glutamine, and serine. Correspondence analysis thus reveals various trends in the amino-acid composition of proteomes of different diazotrophs based on their life-style pattern.

4.2 Evolutionary rate analysis:

4.2.1 Investigation of synonymous substitution rate across various genomes:

The ratio of the number of nonsynonymous substitutions per nonsynonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks) (i.e Ka/Ks) reflects the rate of adaptive evolution against the background rate. This parameter has been widely studied in the analysis of adaptive molecular evolution, and is regarded as a general method of measuring the rate of sequence evolution. Various combinations of symbiotic diazotroph, cyanobacteria and few methanogenes genomes were used for pairwise calculation of Ka/Ks using the Nei-Gojobori method. Average pairwise Ka/Ks values for all protein coding genes (PCGs) and nitrogen-fixing genes (NFGs) are provided in the Table 4.2. From the table it is clear that most of the genes in various genome combinations are under purifying selection with Ka/Ks<<1. PCGs along with NFGs of various *Frankia* pairings have highest average Ka/Ks ratio amongst all the diazotrophs under study. The average Ka/Ks ratio for all possible *Bradyrhizobium* pairs is higher than that observed for *Rhizobium* strains. The evolutionary rate of NFGs (in terms of Ka/Ks ratio) amongst *Frankia* stains is significantly different from those observed in other diazotrophs (Mann–Whitney U test, $P < 0.001$). The increased rate of evolution of nitrogen fixing genes (NFGs) in *Frankia* could be the result of great turmoil going in the *Frankia* genomes in terms of expansion and reduction which itself is a consequence of their response to new environments in the context of their symbioses (Normand *et al.*, 2007). On the other hand, the rate of molecular evolution of nitrogen fixing genes in *Chlorobium* and *Nostoc* strains is quite stable with strong purifying selection which indicates rapid and strong elimination of non-synonymous changes in NFGs.

To gain further insight into the molecular evolution of nitrogen fixing

Table 4.2 Average pairwise Ka/Ks values for all protein coding genes (PCGs) and nitrogen-fixing genes (NFGs) in various diazotroph genomes

Pairwise Combination	Gene Group	Ka	Ks	Ka/Ks
ACN-CCI3	PCG	0.15	0.60	0.25
	NFG	0.14	0.40	0.36
ACN-EAN	PCG	0.23	0.65	0.35
	NFG	0.16	0.39	0.39
EAN-CCI3	PCG	0.22	0.73	0.31
	NFG	0.19	0.54	0.34
DG-ACN	PCG	0.24	0.70	0.35
	NFG	0.24	0.55	0.45
DG-CCI3	PCG	0.25	0.79	0.31
	NFG	0.23	0.63	0.38
DG-EAN	PCG	0.25	0.73	0.34
	NFG	0.25	0.56	0.45
BJAP-BORS	PCG	0.19	0.68	0.28
	NFG	0.18	1.03	0.19
BJAP-BTAI	PCG	0.20	0.74	0.27
	NFG	0.18	1.03	0.19
BORS-BTAI	PCG	0.11	0.58	0.18
	NFG	0.12	0.67	0.15
RTRI-RVICIA	PCG	0.09	0.43	0.17
	NFG	0.11	0.30	0.15
RTRI-RCFN	PCG	0.11	0.68	0.15
	NFG	0.12	1.21	0.10
RTRI-RCIAT	PCG	0.11	0.59	0.16
	NFG	0.14	0.89	0.12
RVICIA-RCFN	PCG	0.10	0.62	0.15
	NFG	0.15	0.67	0.14
RVICIA-RCIAT	PCG	0.12	0.67	0.16
	NFG	0.10	0.60	0.15
RCFN-RCIAT	PCG	0.11	0.57	0.18
	NFG	0.03	0.27	0.17
NOSA-NOSP	PCG	0.17	1.17	0.15
	NFG	0.11	1.08	0.10
NOSA-NOS7120	PCG	0.18	1.37	0.14
	NFG	0.13	1.36	0.10
NOS7120-NOSP	PCG	0.16	1.24	0.14
	NFG	0.14	1.09	0.11
CHPH-CHTE	PCG	0.27	2.31	0.13
	NFG	0.17	2.16	0.08
CHPH-CHLO	PCG	0.22	1.78	0.14
	NFG	0.18	2.04	0.08
CHTE-CHLO	PCG	0.26	1.92	0.15
	NFG	0.16	1.59	0.10
CY51142-CY7425	PCG	0.31	2.74	0.13
	NFG	0.24	2.72	0.09
CY51142-CY8802	PCG	0.22	1.35	0.17
	NFG	0.15	1.40	0.12
CY7425-CY8802	PCG	0.30	2.71	0.12
	NFG	0.25	1.80	0.11
MMAZ-MACE	PCG	0.12	0.92	0.14
	NFG	0.15	1.10	0.12
MMAZ-MBAK	PCG	0.15	1.36	0.12
	NFG	0.12	1.26	0.11
MBAK-MACE	PCG	0.14	1.23	0.13
	NFG	0.09	1.07	0.09
MMARI-MVAN	PCG	0.15	1.08	0.15
	NFG	0.14	1.01	0.13
MMARI-MAEO	PCG	0.26	1.60	0.18
	NFG	0.27	1.58	0.17
MVAN-MAEO	PCG	0.27	1.52	0.19
	NFG	0.28	1.63	0.19

machinery, the rate of nucleotide substitution for *nif*HDK genes which code for the nitrogenase enzyme were considered individually. The exact pairwise Ka/Ks values for the three genes in various pairings of diazotroph are provided in Table 4.3. From the table it is clear that purifying selection plays a role in shaping the evolution of these genes. Amongst the *nif* genes, *nif*H tend to be more slowly evolving than others, indicating even stronger stabilizing selection. The no significant difference was observed in pairwise Ka/Ks value for *nif*H gene in various possible grouping of diazotrophs. The highest Ka/Ks value (0.262) for *nif*H was observed for coupling of two *Methanococcus* strains. In comparison, Ka/Ks value for *nif*D (coding the α subunit of nitrogenase) gene show reasonable variation with values ranging from 0.034 to 0.451. The Ka/Ks value for *nif*D in combination involving *Nostoc punctiforme* could not be calculated because *nif*D in *Nostoc punctiforme* has insertion element that pops out in certain growth condition and the protein is expressed only after removal of insertion element

Table 4.3 The value of pairwise Ka/Ks ratio for *nif*HDK genes in various diazotrophs

	***nif*H**	***nif*D**	***nif*K**
ACN-CCI3	0.113	0.177	0.408
ACN-EAN	0.141	0.219	0.377
EAN-CCI3	0.110	0.215	0.201
DG-ACN	0.216	0.407	0.734
DG-CCI3	0.218	0.340	0.570
DG-EAN	0.207	0.451	0.523
BJAP-BORS	0.046	0.052	0.093
BJAP-BTAI	0.049	0.059	0.089
BORS-BTAI	0.070	0.034	0.036
RTRI-RVICIA	0.054	0.068	0.062
RTRI-RCFN	0.062	0.079	0.080
RTRI-RCIAT	0.052	0.074	0.068
RVICIA-RCFN	0.049	0.078	0.054
RVICIA-RCIAT	0.051	0.079	0.052
RCFN-RCIAT	0.055	0.077	0.082
NOSA-NOSP	0.083	-	0.100
NOSA-NOS7120	0.116	0.144	0.191
NOS7120-NOSP	0.149	-	0.183
CHPH-CHTE	0.028	0.068	0.098
CHPH-CHLO	0.039	0.098	0.080
CHTE-CHLO	0.024	0.087	0.086
CY51142-CY7425	0.061	0.049	0.029
CY51142-CY8802	0.198	0.157	0.140
CY7425-CY8802	0.084	0.069	0.051
MMAZ-MACE	0.035	0.055	0.062
MMAZ-MBAK	0.032	0.050	0.041
MBAK-MACE	0.128	0.051	0.046
MMARI-MVAN	0.048	0.094	0.344
MMARI-MAEO	0.262	0.284	0.365
MVAN-MAEO	0.174	0.206	0.368

(Meeks *et al.*, 2001). A much relaxed purifying selection is observed in case of *nif*K (coding the β subunit of nitrogenase) gene in parings involving *Frankia* strains. Similarly in *Methanoccocus* strains the intensity of purifying selection was significantly relaxed for *nif*K gene. In all of the possible pairwise comparisons of *Rhizobium, Bradyrhizobium, Chlorobium* and *Methanosarcina strains,* more or less similar Ka/Ks values were obtained for *nif*DHK genes. This result indicates that the evolutionary rate of nitrogenase coding genes is quite stable among these strains.

4.2.2 Sliding window analysis of nifHDK gene:

Finer scrutiny of nucleotide diversity in *nif*HDK genes were carried out using sliding window analysis of pairwise Ka/Ks. For the analysis of long genes like that of *nif*HDK, Ka/Ks was measured by taking smaller windows of only a fraction the size of the gene, and computing the ratio for just the window. By sliding this window along the sequence, idea can be gained regarding the range of values of Ka/Ks, and therefore of selection, across the gene. This helps in detection of patches of diversifying selection. Although the general trend of the Ka/Ks ratio is less than 1, there may be some regions where the ratio is greater than one, indicating that these regions are likely to be under positive selection. These segments usually display high amino acid variability. Figures 4.4-4.6 represents the pairwise sliding window analysis of nitrogenase genes from *Frankia* genomes. Rest of the relative figures from other studied diazotroph can be found in Supplementary Figure SF 2.1-2.3.

Analysis for *nif*H gene: Pairwise sliding window analysis of *nif*H genes from various *Frankia* strains (Figure 4.4) revealed that purifying selection dominates the entire length of gene. Although a relaxed purifying selection can be seen in the C-terminal (right) end in all the pairwise combinations. Out of all the possible pairwise comparison, those involving the *Frankia* DG show a similar pattern of distribution of Ka/Ks values while rest of the combination show a slightly different pattern. In *nif*H genes of *Rhizobium* strains (Supplementary Figure SF3.1A), it is seen that extensive purifying selection (Ka/Ks<<1) dominates the intact length of the gene in all the pairwise permutation with average value close to 0.05. The

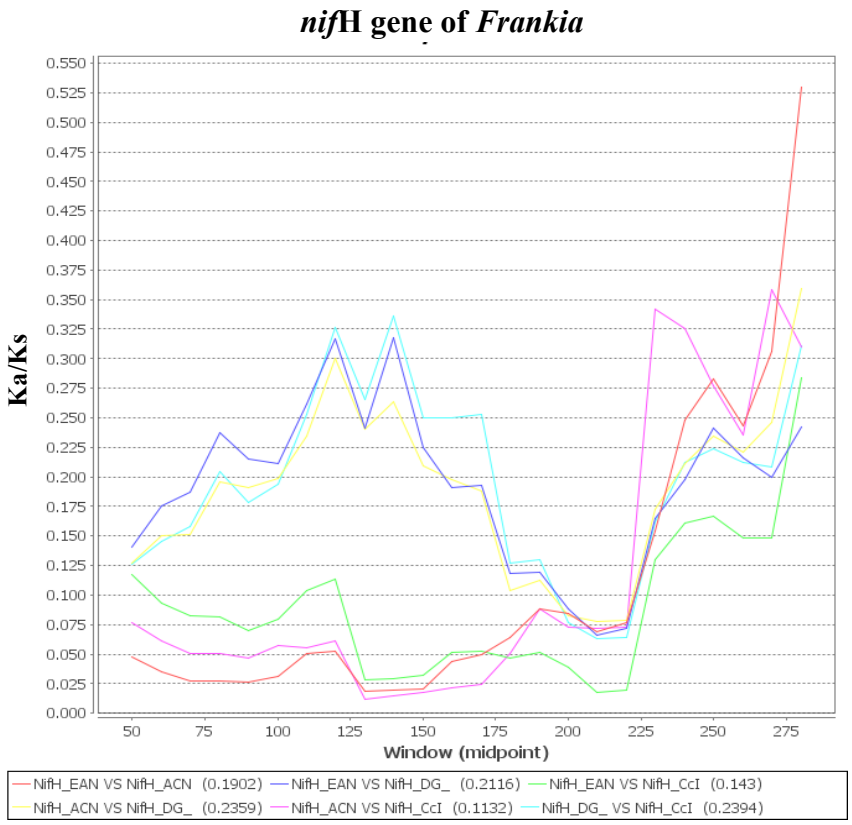


Figure 4.4 Sliding window analysis of pairwise Ka/Ks in *nif*H gene in *Frankia* . All pairwise comparisons were performed using a window 100 and jump of 10.

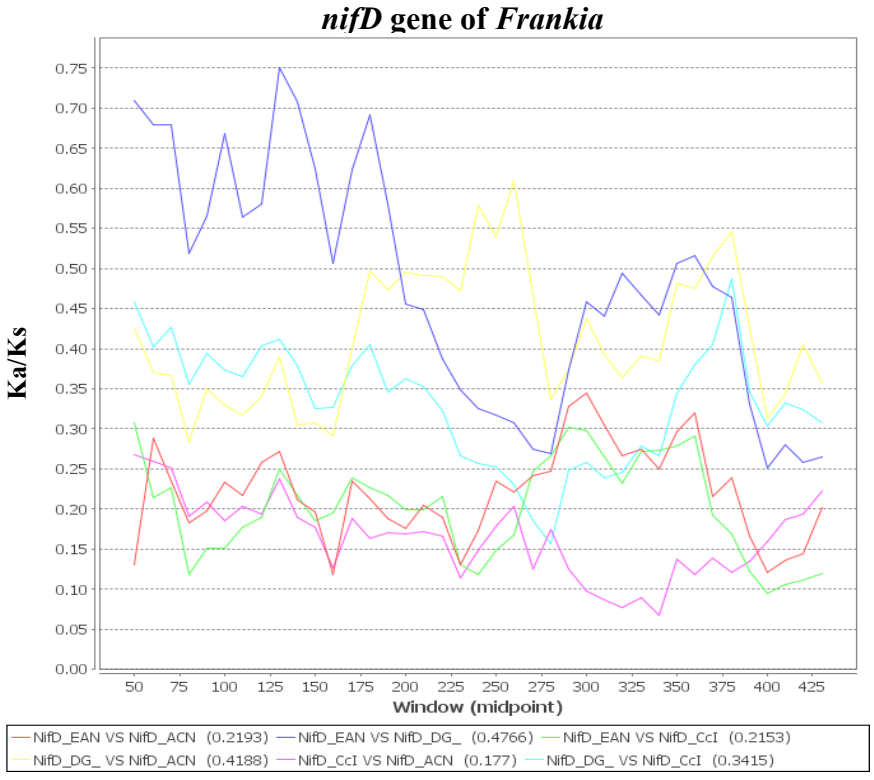


Figure 4.5 Sliding window analysis of pairwise Ka/Ks in *nif*D gene in *Frankia* . All pairwise comparisons were performed using a window 100 and jump of 10.

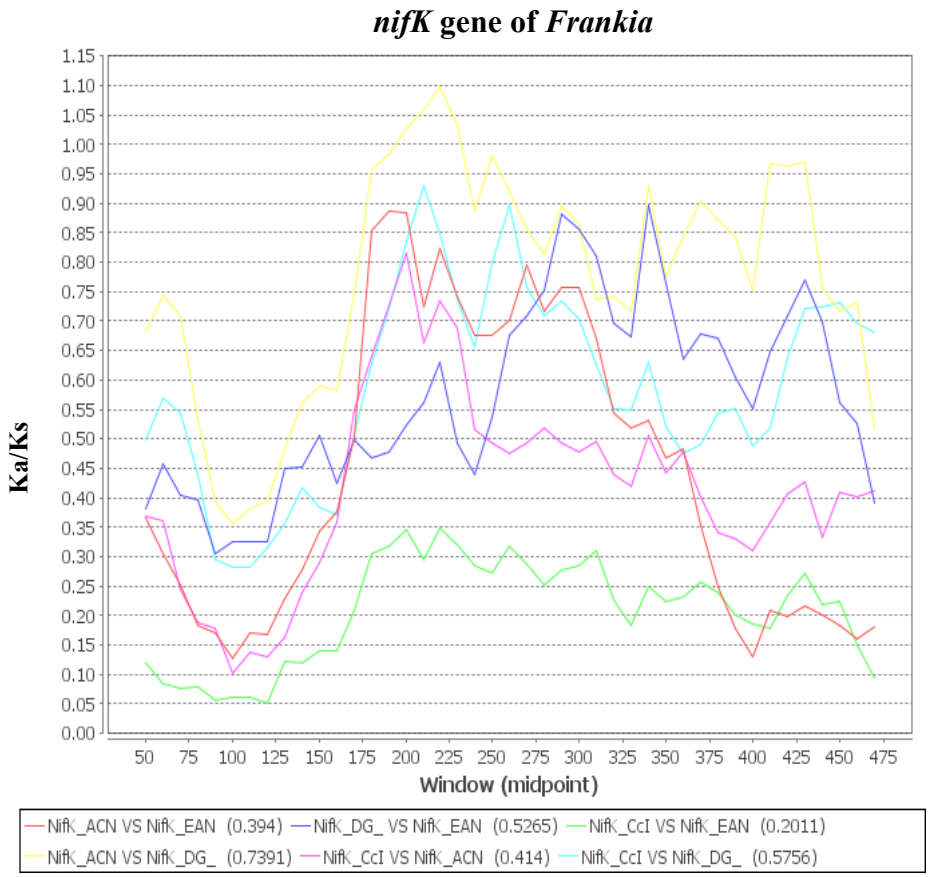


Figure 4.6 Sliding window analysis of pairwise Ka/Ks in *nif*K gene in *Frankia*. All pairwise comparisons were performed using a window 100 and jump of 10.

*nif*H genes of other diazotrophs like *Chlorobium*, *Cyanothece, Methanococcus, Methanosarcina* and *Bradyrhizobium* also experience similar purifying selection pressure throughout the gene. These sliding window analysis results indicate that the evolutionary rate of this gene is quite stable among various diazotrophs.

Analysis for *nif*D genes: Sliding window analysis of *nif*D genes showed that all the pairwise combination of *Frankia* strains experiences purifying selection with Ka/Ks often remaining below 1 in most part of the gene. But relaxed purifying selection is noticed in the region 125-150 in for the *Frankia* EAN1pec/DG combination (Figure 4.5). These regions correspond to amino acid residues involved in alpha and beta subunit interaction in nitrogenase enzyme. In other studied diazotrophs (see Supplementary Figure SF 3.2), *nif*D gene show regions of higher divergence mixed with more stable regions. However, the average ratio for the entire gene is always less than 1 in all the cases. Thus there is seems to be a tendency for elimination of non-synonymous changes in this protein-coding gene.

Analysis for *nif*K gene: Computing the Ka/Ks in sliding window across *nif*K gene for *Frankia* strains revealed that the rate of non-synonymous substitutions is greater than the rate of synonymous substitutions in some regions of *nif*K gene which is a clear indication of positive selection (Figure 4.6). Particularly the conserved domain involving P-cluster binding residues in *Frankia* ACN14a/DG combination are found to under positive selection pressure. The P-cluster is particularly important in electron transport mechanism of nitrogenase system (Chan *et al.*, 1993). In other pairwise combinations of *Frankia*, even though the Ka/Ks ratio is below 1 but a very relaxed purifying selection is noticed in this domain indicating high amino acid variability. Least variation in Ka/Ks is noticed for the Frankia CcI3/EAN1pe comparison. In diazotroph like *Rhizobium* and *Chlorobium,* the Ka/Ks ratio remains below 1 in most of the region of the *nif*K gene. Across the *nif*K gene of *Methanococcus*, the rate of non-synonymous substitutions has been found to be quite higher in most of the combinations, although the average Ka/Ks over the entire gene is below 1 (see Supplementary Figure SF3.3).

4.3 Genomic comparisons:

4.3.1 Comparative genomics of Rhizobiales :

4.3.1.1 Gene Family clustering and construction of Pan- and Core Genome:

The genomes of 15 nitrogen fixing strains of the order Rhizobiales were compared. The genes were assigned into gene families based on sequence similarity using BLASTP program. BLAST results were analyzed to construct a pan-genome and a core genome. A plot of the evolution of the pan- and core genome as more and more genomes are considered is seen in Figure 4.7. The plot shows that the pan -genome (blue line) increases with the addition of a second genome while the core (red line) genome decreases considerably with the addition of the 2nd genome and further dropping slowly with the addition of the 3rd and 4th genome until it reaches what appears like a stable level after 5-6 genomes. This is because every genome can add many novel (and frequently different) genes to the pan-genome but only decreases the core genome with a few genes that are absent in that particular strain but that were conserved in the previously analyzed genomes. Finally when all the

Rhizobiales (15 genomes) have been considered, there are 1014 gene families in the core-genome and 31062 gene families in the pan-genome. Details of the pan-genome and new gene families being added are also provided in Table 4.4.

In the step going from *Azorhizobium caulinodans* to *Bradyrhizobium* sp. BTAi, there is a steep decline in the number of core genes (almost by 41%), most likely because the stem nodulating photosynthetic *Bradyrhizobium* sp. BTAi genome lacks many vital nodulating genes (Nod genes) (Giraud *et al.*, 2007) . Due to greater conserved gene families, the core-genome remains stable after 5th genome onwards while the pan-genome continues to increase, but at a relatively slow rate with fewer new genes added to the pangenomes. Thus it seems that members of Rhizobiales exhibits what has tentatively been called a "closed" pangenome structure (Tettelin *et al.*, 2008). This is in contrast to the *E. coli* which clearly displays an open pan-genome structure (Rasko *et al.*, 2008), but congruent with pathogens such as *Salmonella* (Jacobsen *et al.*, 2011), *Listeria* (Deng *et al.*, 2010), or *Campylobacter jejuni* (Friis *et al.*, 2010).

The pan and core genomes of pairwise genome comparisons were also determined to establish the percentage identity for each combination. This identity was expressed as the pairwise

Table 4.4 Details of the pan-genome and core-genome of the Rhizobiales members

Genomes being added	Total genes	New genes	New families	Pan genome	Core genome
Azorhizobium caulinodans ORS571	4717	4717	4452	4452	4452
Bradyrhizobium_BTAi1	7741	5504	5063	9485	1846
Bradyrhizobium_japonicum_USDA110	8317	3529	3326	12713	1691
Bradyrhizobium_ORS278	6752	1193	1163	13812	1642
Mesorhizobium_ciceri	6470	4255	4066	17832	1168
Mesorhizobium_loti	7272	2184	2059	19802	1133
Mesorhizobium opportunistum_WSM2075	6674	931	903	20646	1132
Rhizobium_etli_CFN42	6030	2685	2548	23125	1059
Rhizobium_etli_CIAT652	6056	1041	1032	24083	1058
Rhizobium leguminosarum_bv_trifolii_WSM1325	7232	1542	1453	25454	1052
Rhizobium _leguminosarum_bv_viciae	7276	1010	976	26380	1043
Rhizobium_NGR234	6376	1848	1730	27992	1020
Sinorhizobium_medicae_WSM419	6518	1511	1398	29302	1015
Sinorhizobium_meliloti_AK83	6954	1193	1128	30341	1015
Sinorhizobium meliloti SM11	7093	775	756	31062	1014

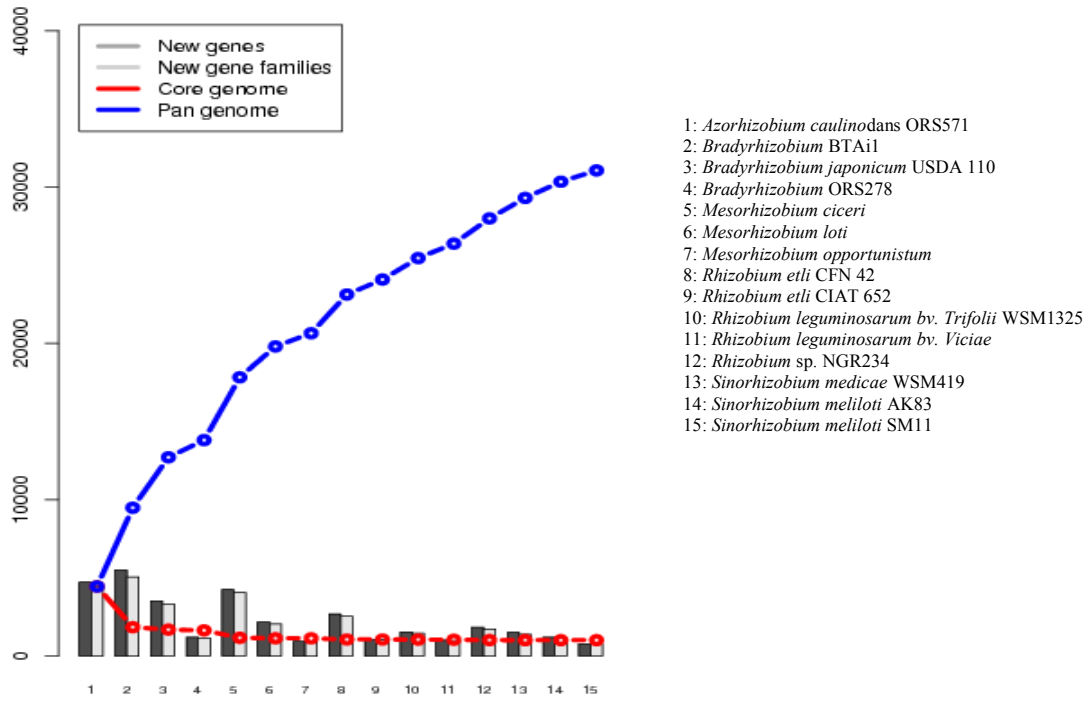


Figure 4.7 The pan-genome (blue line) and core genome (red line) plot for Rhizobiales genomes. The number of discovered novel genes (dark bars) and novel gene families (light-grey bars) are also shown for each added genome

core genome divided by its pangenome and was visualized by color intensity in a BLAST Matrix. Figure 4.8 shows the BLAST Matrix for the Rhizobiales genomes. The darkest green, indicative of the highest fraction of genes found similar between two genomes. The highest similarity recorded is 71.6%, between two *Sinorhizobium meliloti* strains, and the lowest is 11.0%, between *Bradyrhizobium japonicum* USDA110 and *Sinorhizobium meliloti* SM11. All the *Rhizobium* genomes, except for the *Rhizobium* NGR234, shares 50.1% to 63.7% of their genes. Amongst the *Mesorhizobium* genomes, *M. opportunistum* WSM2075 and *M. ciceri* share a large fraction (63.2%) of genes between them.. The highest identity (43.1 %) between the genomes of two genera is observed in case of *Rhizobium* NGR234 and *Sinorhizobium medicae* WSM419, probably lending support to synonymy of *Sinorhizobium* and *Rhizobium* as suggested in earlier works (Chen *et al.*, 1988; Jarvis *et al.*, 1992). The highest homology (6.9%) within the organism's own proteome is observed for the *Bradyrhizobium* BTAi1 genome indicating the possibility of gene duplication within this genome.

The conservation of genes across the Rhizobiales genomes was assessed and reported irrespective of their function, but functional information is essential

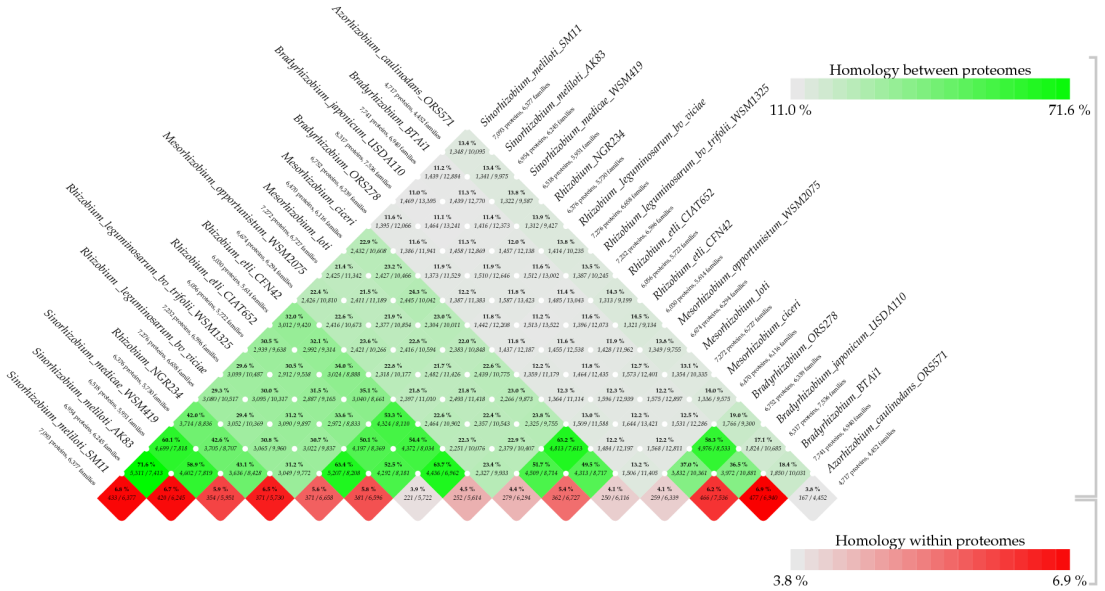


Figure 4.8 BLAST Matrix for the Rhizobiales genomes. To the side, the total number of protein genes and gene families are listed for each genome. In the matrix cells, the shared protein genes are given as a percentage, based on the ratio of the core genome and pan-genome of each pair (for a figure with higher resolution see Supplementary Figure SFHR 4.8)

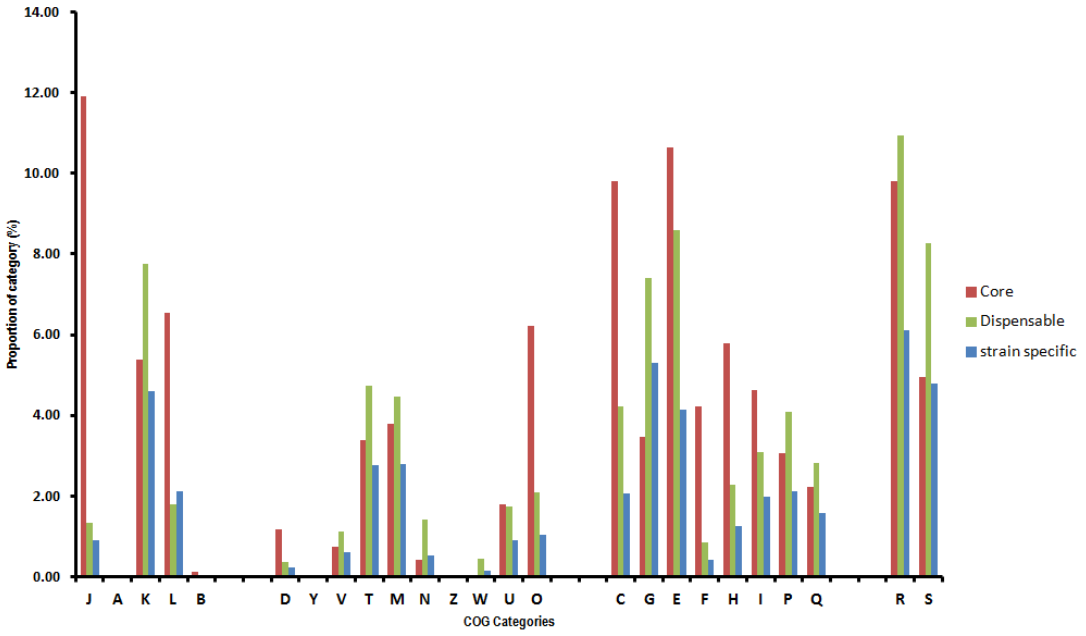


Figure 4.9 COG statistics for the genes found in core, dispensable and strain-specific set of Rhizobiales genomes

for a biological interpretation. Therefore the functional categories for the core and accessory genomes for Rhizobiales strains were inferred through their placement into different COG Categories and the results were summarized in Figure 4.9. The core and accessory genomes together constitute the pangenome. The accessory genome has further been

broken into two set- dispensable and strain specific. The dispensable set include genes are those found in more than one strain, but not in all strains while strain specific set include genes unique to a single strain. For the core genes a greater focus can be seen on intracellular activities exemplified directly by the higher fraction of the core genome being associated with translation, ribosomal structure and biogenesis, transcription, replication, recombination and repair, cell wall/ membrane/envelope biogenesis, amino acid transport and energy production. For instance, within COG Functional group 'information storage and processing', class J (translation, ribosomal structure and biogenesis) is enriched in the core genome, at the expense of classes K and L (transcription and replication, respectively). Of interest is also the shift within the group 'metabolism' between classes E and G (for amino acid and carbohydrate transport/ metabolism, respectively). The results indicate that the gene content for metabolism of amino acids is more conserved than that for carbohydrates, in the Rhizobiales genomes. This is apparently due to the nitrogen-fixing properties of these genomes. Lastly, enrichment in the core genome of class O, for post-translational modification and chaperones, is evident within the group 'cellular processes and signaling'. On the other hand, dispensable clusters and strain-specific clusters are well represented in transcription (class K), signal transduction (class T), cell wall/ membrane/envelope biogenesis (class M) and inorganic ion transport categories (class P). Thus the genes belonging to these clusters show a higher degree of involvement in extracellular activities like transporting molecules in or out of the cell. It is also noticed that strain-specific clusters are rich in carbohydrate transport and metabolism, which may be related to their different living niche.

4.3.1.2 Organization of nitrogen fixation related genes across the genome:

Protein comparisons along with gene family clustering have allowed the identification of a variety of similarities and differences in several groups of genes, including those involved in nodulation and nitrogen fixation. The important nitrogen fixation related genes found in most members of Rhizobiales includes the *nif* genes, *fix* genes along with

nodulation genes. In many rhizobial species nodulation (*nod*, *nol* and *noe*), *nif* and *fix*, as well as many secretion related genes, are found in a symbiotic plasmid, while in species or strains without plasmids, the genes are located in laterally transferrable genomic islands, also denominated as symbiotic islands. These symbiotic regions of nitrogen fixing Rhizobiales genomes have been found to be largely mosaic structures that have been frequently altered by recombination, horizontal transfer and transposition events (Gonzalez *et al.*, 2003).

The results of gene family clustering were utilized for the extraction of all *nod* protein ortholog clusters (Table 4.5). In total, twenty eight different Nod protein ortholog clusters amongst the fifteen species were found. All fifteen species contained *nod*E, *nod*G, *nod*M, *nod*P, *nod*Q and *nod*V genes. One of the Nod orthologs found in all species, *nod*G, is a large gene family cluster. It is a cation protein exporter, involved in the secretion of the finished Nod factor into the environment (Oliveira *et al.*, 2010). Two *nod* genes were found exclusively in one strain: *nod*R in *R. leguminosarum* bv. Trifolii WSM1325 and *nod*O in *R. leguminosarum* bv. viciae 3804. The function of *nod*R is unknown; however, it is suspected that NodR might contribute to the superior host nodulation efficiency of WSM1325, compared to WSM2304 (Yates *et al.*, 2011). In relation to NodO, it catalyzes the addition of carbamoyl to the Nod factor backbone (Downie, 2011), and it is also a calcium-binding protein that promotes infection thread development in root hairs (Walker *et al.*, 2000). The stem nodulating symbiont *Bradyrhizobium* ORS278 had the least number of *nod* genes. Surprisingly, the two *Bradyrhizobium* genomes, *Bradyrhizobium* sp. BTAi1 (btai) and *Bradyrhizobium* sp. ORS 278 (bors), have no *nod*A, *nod*B or *nod*C genes and thus their signaling mechanisms with the host plant are still not fully understood (Bonaldi *et al.*, 2010).

Two other important gene families that are involved in Nod factor synthesis and regulation includes *nol* and *noe* genes. As seen Supplementary Table ST3, protein NolG is ubiquitous and present in all strains under the study. The NolG cluster contains many orthologs (around 130), many belonging to the very broad COG category of cation/multidrug efflux pump. Therefore it is possible that several orthologs represent only non-

Table 4.5 Nod protein orthologs in Rhizobiales genomes

Coding gene	azo	bjap	btai	bors	mciceri	mloti	mop	rcfn	rciat	rtrifolii	rviciae	rngr	smed	smelAK83	smelSM11
nodA															
nodB															
nodC															
nodD															
nodE															
nodF															
nodG															
nodH															
nodI															
nodJ															
nodL															
nodM															
nodN															
nodO															
nodP															
nodQ															
nodR															
nodS															
nodT															
nodT															
nodT															
nodU															
nodV															
nodW															
nodW															
nodX															
nodY															
nodZ															

Ortholog in Azorhizobium Bradyrhizobium Mesorhizobium Rhizobium Sinorhizobium

Grey cells indicate absence of gene

nodulation cation pumps. Many differences were noticed in noe clusters as well (Supplementary Table ST4). The maximum number of genes in this category was found in *Bradyrhizobium japonicum* USDA 110 (bjap). *noe*D gene is unique for bjap does not have a ortholog in any other strains. *Sinorhizobium* species have only *noe*A and *noe*B, and these genes were thought to be specific for *Medicago* nodulation (Ardourel *et al.*, 1995).

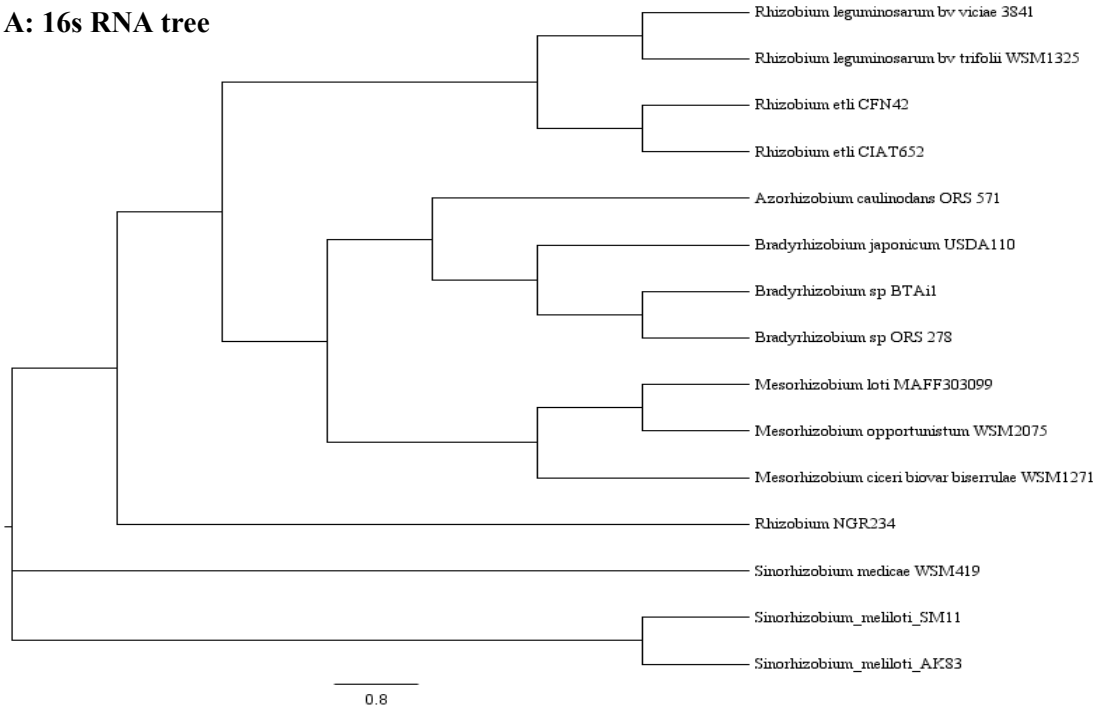
The fixation of nitrogen via nitrogenase requires an anaerobic or micro-aerobic bacteroid environment and the *fix* gene family is involved in the regulation and metabolism of oxygen in this circumstance. The *fix* gene family is commonly found in three core operon structures: *fix*ABCX, *fix*GHIS and *fix*NOPQ. All fifteen symbiotic species have the complete set of genes of the three operons (Supplementary Table ST5). Each different species, except for *Azorhizobium caulinodans* ORS 571, has a unique *fix*S ortholog cluster, and *Mesorhizobium loti* MAFF303099 has two. However, as the FixS protein is only fifty five amino acids in size, this result could also be related to a limitation of the BLAST algorithm. The stem nodulator, btai, has no *fix*Q ortholog, which codes for cytochrome oxidase. This diversity of both FixQ and FixS proteins may suggest a diverse range of cytochrome cbb3 oxidase complexes for each species that could result from differences in host nodulation strategies.

The *nif* family codes for the MoFe-dependant nitrogenase complex, the enzyme required for the catalysis for the nitrogen fixing process. The *nif*BHDKENX operon which codes for the main elements of nitrogenase complex is more or less present in all the fifteen symbiont (See Supplementary Table ST6). However, the *nif*X gene is totally absent in the two *Rhizobium leguminosrium* strains (rtri and rvic). *nif*Q gene which is required for Mo-incorporation into the nitrogenase complex is lacking in *R. leguminosarum and Sinorhizobium* strains. Consequently, these species should use other means to incorporate Mo into the nitrogenase complex. Orthologs of *nif*U gene product which is required for the full activation of the metalloenzyme nitrogenase and nif regulator genes like *nif*A and *nif*R are presents in all the rhizobiales strains under consideration.

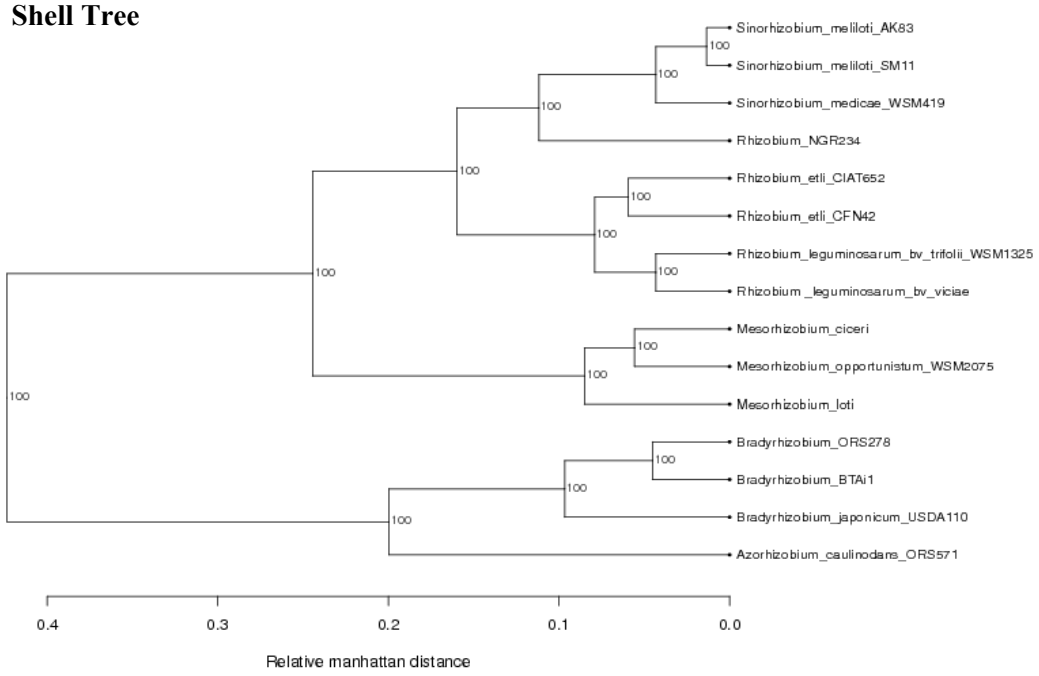
4.3.1.3 Clustering of Rhizobiales genomes:

The phylogeny and taxonomy of nitrogen fixing Rhizobiales is in a state of flux. Therefore in this genomic comparison several methods were utilized to highlight different aspects of the phylogeny of the selected strains. For bacterial genomes, the 16S rRNA genes are often used for identification of bacteria at the genus and species level. The phylogenetic tree presented in Figure 4.10 A , is based on 16S rRNA gene sequences extracted from

A: 16s RNA tree



B: Shell Tree



Continued to next page

Continued from previous page

C: Cloud Tree

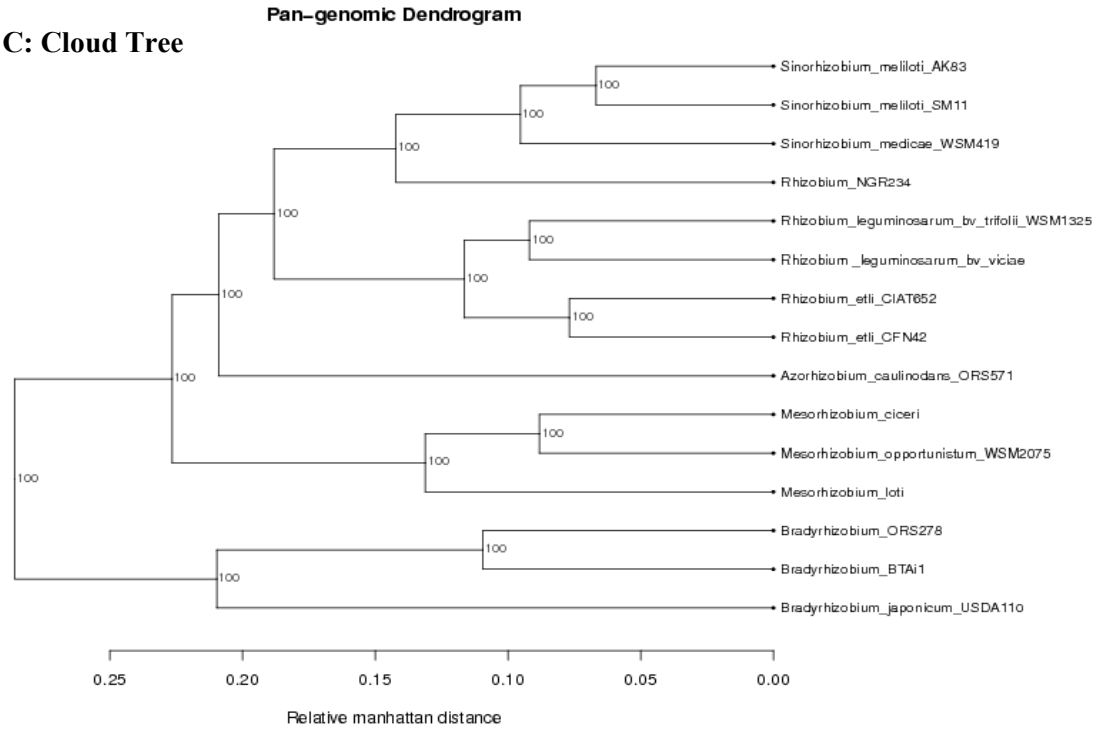


Figure 4.10 (A) 16S rRNA tree ; (B) Shell-type pan genome tree ; (C) Cloud-type pan genome tree for fifteen Rhizobiales genomes

15 genomes using the HMM-based RNAmmer program. The 'classical' relationship between *Bradyrhizobium*, *Rhizobium*, *Sinorhizobium* and *Mesorhizobium* is clearly shown in the 16S rRNA phylogeny. The four major groups are delineated from each other forming four major clusters, with *Azorhizobium* present as a separate branch. All strains were clustered within the expected genera except for *Rhizobium* NGR234, which is placed quite far apart from the *Rhizobium* clade and closer to the *Sinorhizobium* clade. This result again justifies the inclusion of this strain under *Sinorhizobium* Genera.

A pan-genome hierarchical clustering tree based on the gene family content was also generated (Figure 4.10 B &C). The tree was generated based on the presence or absence of specific gene families across the genomes of the members of order rhizobiales. The pan-genome tree is an alternative presentation of the data previously shown in Blast matrix figure but emphasizes on inter-species relationships. As expected, genomes reporting a high similarity score in the BLAST matrix, cluster together on the pan-genome tree. Two dendrograms (Shell and Cloud) constructed from the overall genomic content of Rhizobiales

strains by weighing the presence of non -core gene families according to two different schemes. The left panel in Figure 4.10 B shows a dendrogram known as 'Shell' tree where gene families are weighted higher the more strains they are present in, while the right panel (Figure 4.10 C) displays a 'Cloud' tree made from weighing gene families higher the fewer genomes they are present in (Snipen and Ussery, 2010). Both type of pan genome tree clustering is similar with four major clades representing the four major genera. But the position of the genera *Azorhizobium* is different in both types of trees. *Bradyrhizobium* and *Azorhizobium* show close relation in the shell type pangenome dendrogram emphasizing on genes present in most strains, but when we give more weight to genes present in few strains, this image reverses and the two organisms become far apart. This is likely the result of distinct nodulation functions as well as putative type-IV secretion system genes (Lee *et al.*, 2008) being present in *Azorhizobium* genome which will be much more significant when rare genes are weighted the highest.

4.3.1.4 Comparison of the Codon usage and amino acid usage:

Comparing the amino acid and codon usage between many genomes also usually involves clustering genomes with similar usage. It is therefore useful to compare these numbers by constructing a heatmap using the statistical package R. Figure 4.11 A&B represents a heat maps produced from the codon and amino acid usage information of the 15 Rhizobiales members. A representative 'heatmap' on codon usage aptly explains the differences in compositional bias among the Rhizobiales genomes. Most of the Rhizobiales members have moderate GC Content around 61-67 percent. As expected the genome thus shows a subtle bias for the GC rich codons like GCC, GGC, GCG,CTG etc. This is quite evident from the colour shades of the codon usage heatmap. The amino acid heat map points to the fact that leucine (L), alanine (A), glycine (G) and Valine (V) are the most utilized amino acids of the proteomes under consideration. The dendrograms on the side of the plots show the correlation between codon/ amino acid frequencies. The organism with similar codon/amino acid usage bias tends to cluster together in the dendrogram. The dendrogram based on codon usage and amino acid usage show similar type of clustering with

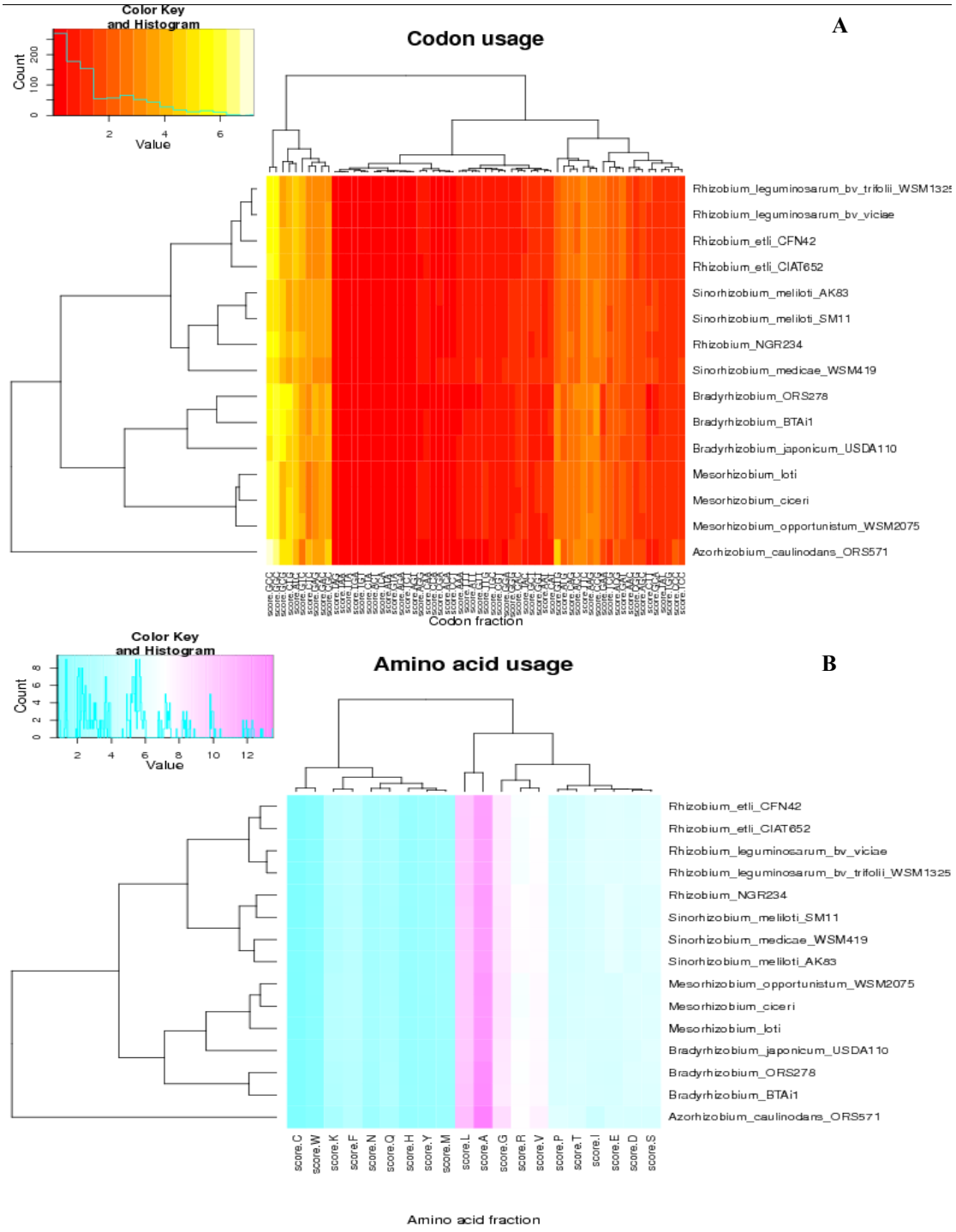


Figure 4.11 Heat map for Rhizobiales genomes based on (A) Codon Usage (B)Amino acid usage. The degree of shading from dark to light color indicates low to high frequency of occurrence of the respective codon/amino acid.

four major clades with *Rhizobium* NGR234 clustering with the *Sinorhizobium* strains rather than *Rhizobium* strains. In both the dendrograms, the *Azorhizobium* tends to remain in an isolated branch. These observations are similar to those observed for Pangenome tree and 16s RNA Tree described previously.

4.3.2 Comparative genomics of

Frankia members:

4.3.2.1 Pan and Core-Genome Construction:

The BLASTP results were utilized for construction of pan and core genome for five strains of *Frankia* and the resulting plot is shown in Figure 4.12. As expected; moving along to the new genomes of *Frankia* members, there is a significant jump for both of pan and core genome curves. By the addition of the accessory genes, the pan genome (blue line) is increased; in contrast core genome (red line) is dropped. After 3rd genome onwards, the core-genome remains almost stable whereas the pan-genome continues to increase steadily. Finally there are 1531 gene families in core genome while the pan-genome constitute of 15409 gene families. Details of the pan-genome and new gene families being added are also provided in Table 4.6

Reciprocal pairwise whole-genome comparisons were also done and overviews of the results were provided in a BLAST matrix, as shown in Figure 4.13. The stronger a matrix cell is coloured, the more similarity was detected between the gene content of two genomes. As can be seen in the triangle, all *Frankia* genomes under consideration are quiet distant with similarity ranging between 18.7% and 31.5% for any given pair of genomes. The highest similarity (31.5%) is noticed for the genomes *Frankia* ACN14a and *Frankia* CcI3. The strains *Frankia* Ean1pec and *Frankia* CcI3 have similarity of around 27.4% in their genome. Therefore it seems that genome Frankia CcI3 is placed somewhere between the genomes of *Frankia* EAN1pec and *Frankia* ACN14a. Strangely, the strains *Frankia* Ean1pec and *Frankia* Eul1c which have same host plant *Elaeagnus* shares merely 26% of their genes. Highest similarity within the same genome is noticed for *Frankia* Ean1pec (8.8%) followed by *Frankia* Eul1c (7.6%). It seems that these genomes have undergone considerable gene duplication. The results obtained in the blast matrix lend support to the point raised in the earlier works of Normand

Table 4.6 Details of the pan-genome and core-genome of the *Frankia* genomes

Genomes being added	total genes	new genes	new families	pan genome	core genome
Frankia Acn14a	6723	6723	6271	6271	6271
Frankia CcI3a	4548	1762	1560	7812	2591
Frankia DG	4728	2303	2055	9827	1748
Frankia Ean1pec	7191	3277	2857	12564	1675
Frankia Eul1c	7205	3126	2927	15409	1531

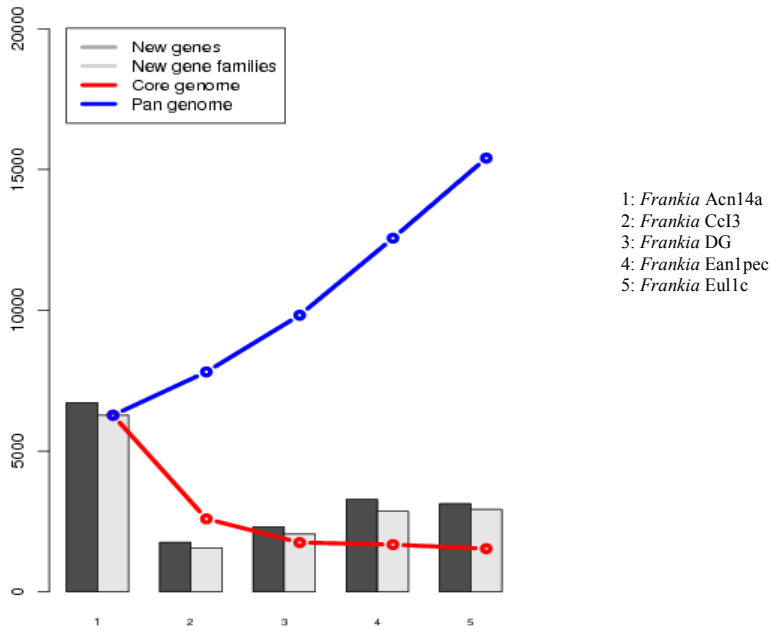


Figure 4.12 The pan-genome (blue line) and core genome (red line) plot for *Frankia* genomes. The number of discovered novel genes (dark bars) and novel gene families (light-grey bars) are also shown for each added genome

et al., (2007) that variation in the genome characteristics of *Frankia* strains is in concert with the differences in host range and host plant biogeography.

Functional characterization for the core and accessory genomes for *Frankia* were inferred through the assignment of the genes into COG categories. The accessory genomes further comprises of the dispensable and strain specific set. Figure 4.14 shows the distribution of the conserved genes in various COG categories. As seen in case of the core genome of members of Rhizobiales, large number of the genes are associated with COG Functional group of "information storage and processing" particularly in classes J (Translation, ribosomal structure and biogenesis), K (Transcription) and L (Replication). Within the COG group of "metabolism", highest percentage of core genome genes are found to be involved amino acid transport (E) followed by coenzyme transport (H) and energy production (C). Interestingly, Carbohydrate transport and metabolism (G) doesnot seems to be priority for core genome genes. The dispensable gene set is well represented in COG class K (transcription) and C (energy production) and I (Lipid transport). The COG class I seems to have large number of dispensable set genes which is markedly different from that observed for Core genome genes and

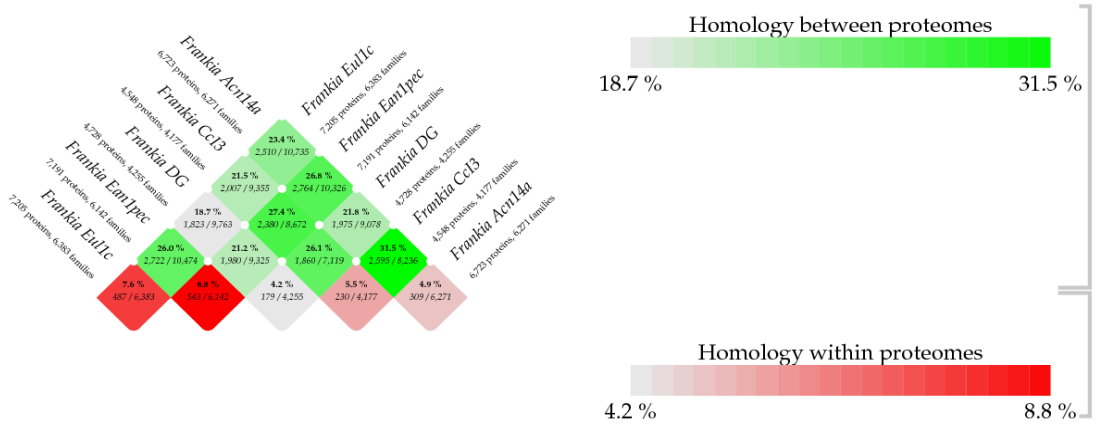


Figure 4.13 BLAST Matrix for the *Frankia* genomes To the side, the total number of protein genes and gene families are listed for each genome. In the matrix cells, the shared protein genes are given as a percentage, based on the ratio of the core genome and pan-genome of each pair.

Strain-specific genes. As expected the, large fraction of dispensable set of genes belong to COG class R associated with general function with many of them being the hypothetical proteins. For strain-specific set of genes a greater focus can be seen on COG classes K, I,Q and R. The secondary metabolites biosynthesis (Q) seems to be important biological function for the genes specific for particular strains.

4.3.2.2 Organization of nitrogen fixation related genes across *Frankia*

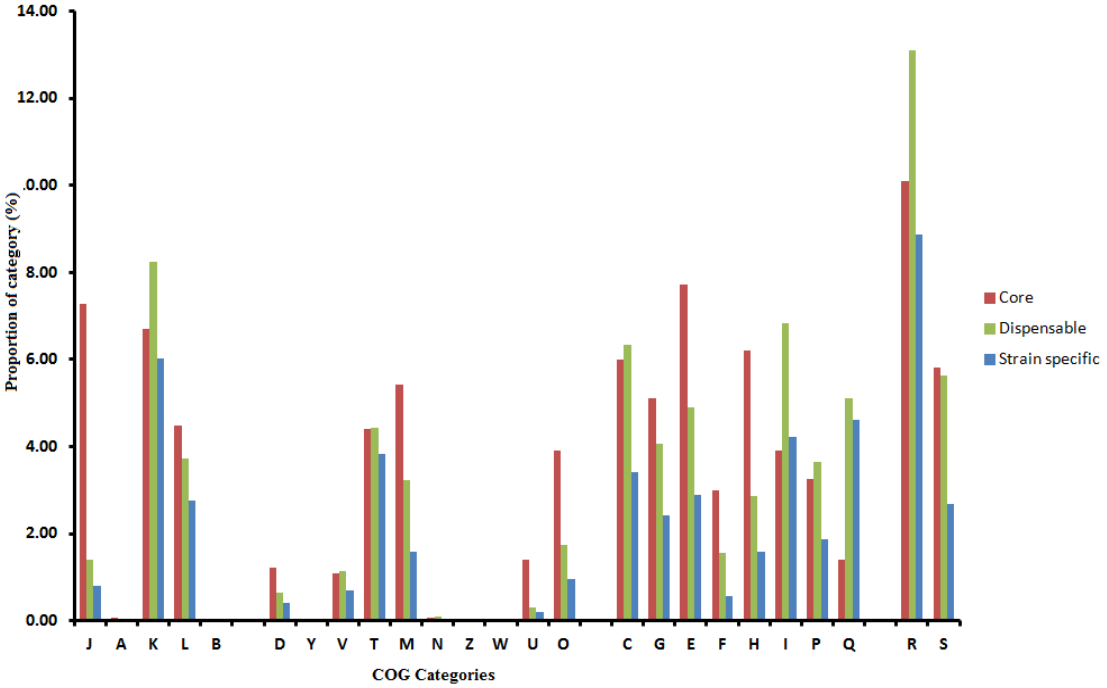


Figure 4.14 COG statistics for the genes found in core,dispensable and strain-specific set of *Frankia* genomes

genomes:

The results of gene family clustering were utilized for the extraction of ortholog clusters of nitrogen-fixation and symbiosis related genes across the *Frankia* genomes (Supplementary Table ST7). The important nitrogen fixation and symbiosis related genes in *Frankia* genomes includes *nif, shc, hup, suf* and *isp* genes (Kucho *et al.*, 2010). Nodulation related genes like *nod,noe* and *nol* is totally missing from these genomes. In total, fifty different Nif proteins are found across the five strains of *Frankia* genomes under observation. Except for *nifU* gene, which involved in FeS biosynthesis, *nif* gene orthologs is totally missing in the genome of *Frankia* Eul1c confirming its existence as an ineffective strain. *Franki*a ACN14a has two ortholog of the gene *nifZ.* Genes like *isp, suf, shc* which are involved in sulfur-iron cluster formation, hopanoid synthesis respectively are universally present in all the strains of *Frankia.* The *hup* genes which are responsible for hydrogen uptake are another vital component of symbiotic machinery of *Frankia.* The *hup* genes show quite diversity in terms of their ortholog. Many of the *hup* genes like *hup*D1, *hup*D2, *hup*L2, *hup*S2 are totally missing in *Frankia* Eul1c. *Frankia* EAN1pec is also missing genes for *hup*D2, *hup*S2.

4.3.2.3 Clustering of *Frankia* genomes:

16S rRNAs are functionally conserved and relatively long, making them ideal for phylogenetic studies. Phylogenetic tree was constructed based on 16S rRNA of *Frankia strains* and are shown in Figure 4.15 A. The phylogenetic tree reveals that the two *Elaeagnus* isolates, EAN1pec (EAN) and EuI1c (EuI) show greater affinity and are grouped together whereas the obligate symbionts like *Frankia* CcI3 and *Frankia* Dg are the two outgroups and the strain *Frankia* ACN14a enjoys a intermediate position in the phylogenetic clustering. This difference in clustering is further substantiated by the differences in the genome size. While *Frankia* EAN1pec and EuIc have large genome sizes (>8 Mbp), *Frankia* Dg and *Frankia* CcI3 have small genome sizes (>5Mbp) and *Frankia* ACN14a has a midway genome size of around 7.4 Mbp.

Two dendrograms constructed from the overall genomic content in Frankia are illustrated in Figure 4.15B&C by weighing the presence of gene families according to two different schemes. One emphasizing the shared shell

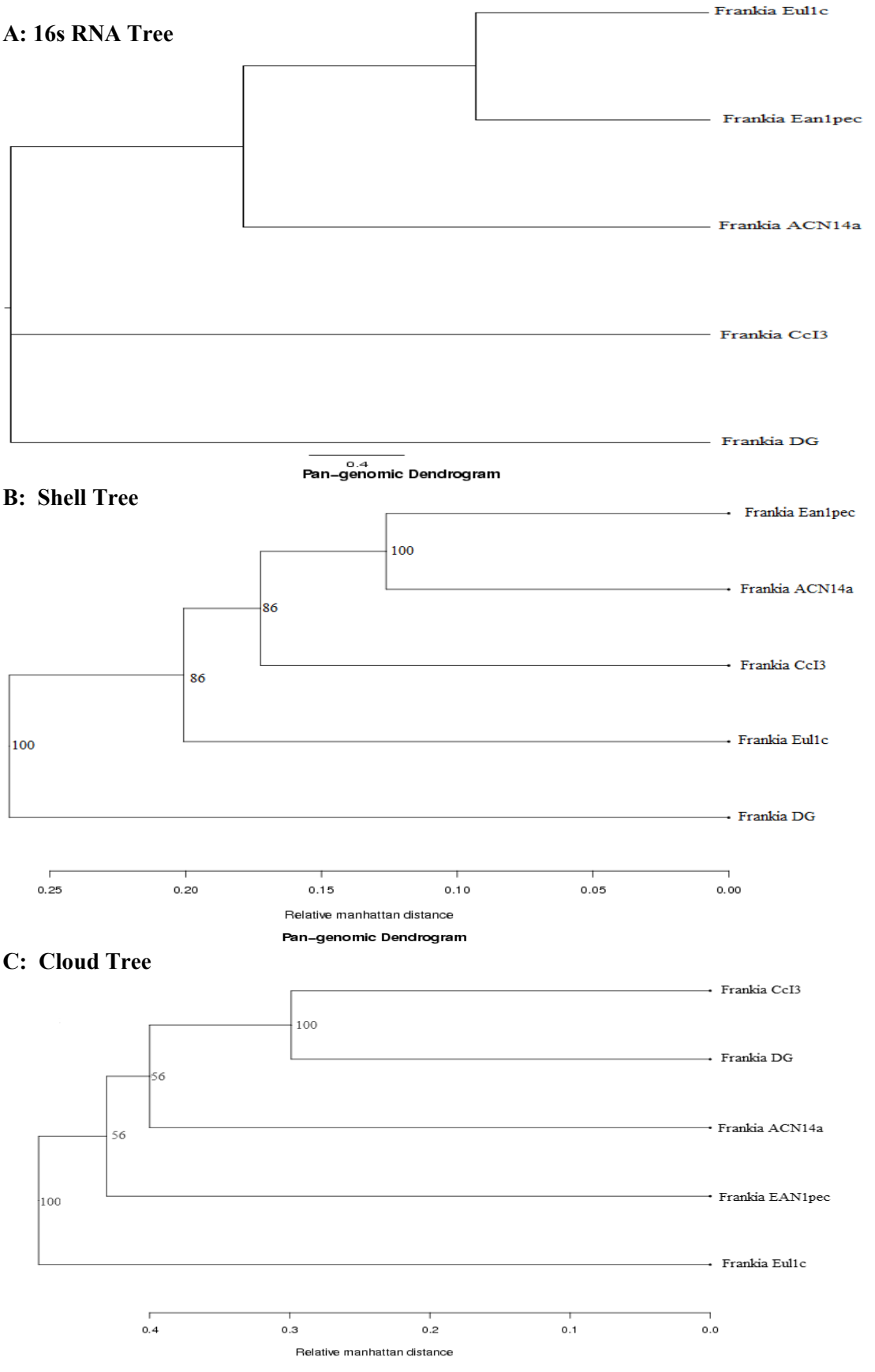


Figure 4.15 (A) 16S rRNA tree ;(B) Shell-type pan genome tree; (C) cloud-type pan genome tree for five *Frankia* genomes

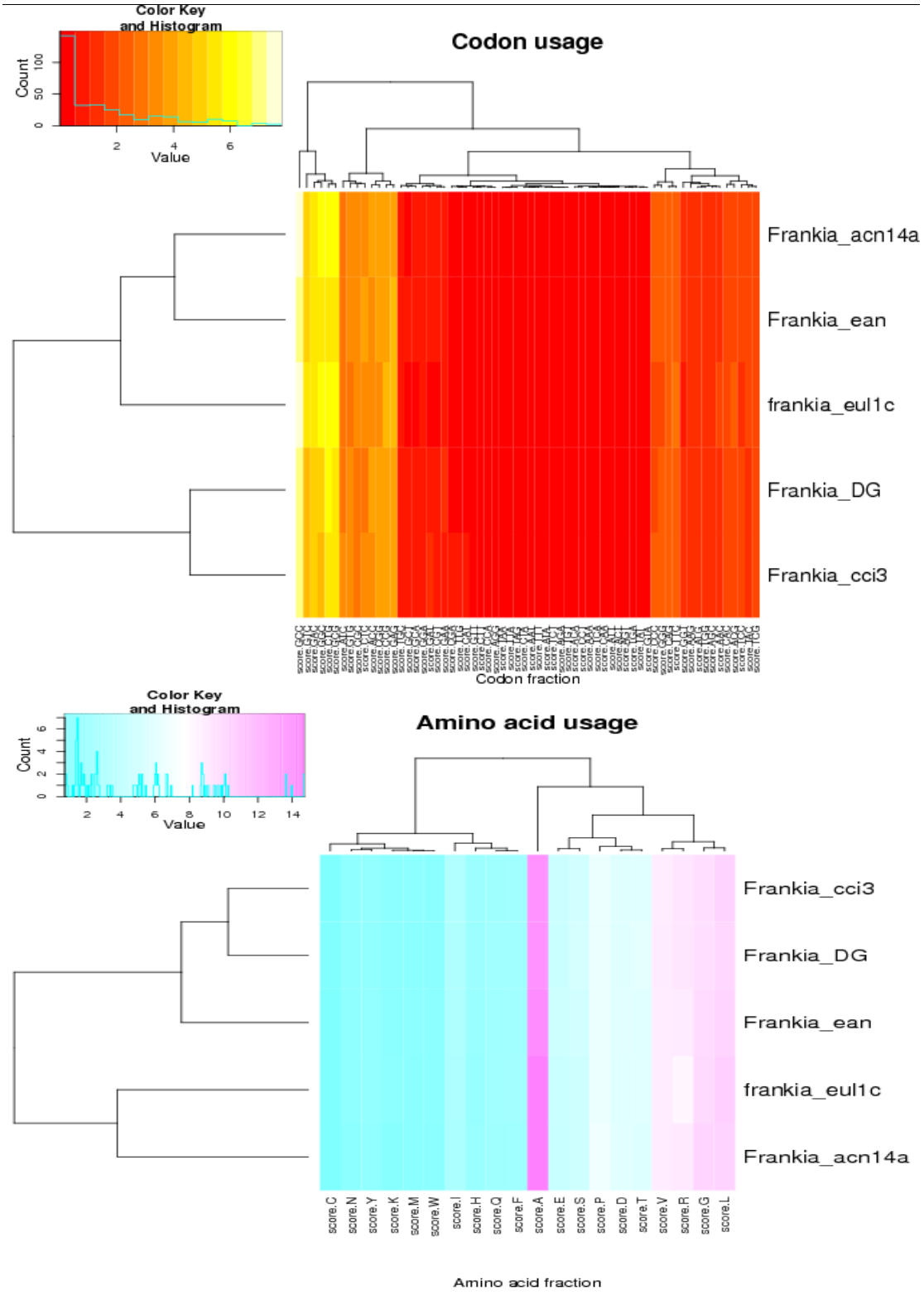


Figure 4.16 Heat map for *Frankia* genomes based on (A) Codon Usage (B) Amino acid usage.

genes (Figure 4.15 B) and the other the less conserved cloud genes (Figure 4.15 C). The shell genes are frequently observed in the pangenome, and differences in shell-gene content most likely reflect an evolution over a longer

time span (Snipen *et al.*, 2009). The shell tree shows that the strain *Frankia* Dg is significantly different from the others. *Frankia* ACN14a and *Frankia* EAN1pec show close proximity in this tree and seems to share many of their genes. The cloud weighted tree, which is based on promiscuous genes, shows *Frankia* CcI3 and *Frankia* Dg are quite similar in terms of gene-content but *Frankia* eul1c and *Frankia* Ean1pec forms two outgroup.

4.3.2.4 Comparison of the Codon usage and amino acid usage:

Heat map produced from the codon and amino acid usage information of the 5 Frankia strains is presented in Figure 4.16 A&B. The overall GC content of the *Frankia* genomes ranges from 70 to73 %. In accordance with the overall high GC content, the codon usage is shifted toward GC-rich codons (Figure 4.16 A) like CTG, GCG, GCC etc. Consequently, GC-coded amino acids like Alanine, Glycine, Leucine and Valine are overrepresented (Figure 4.16 B). Correlation of Codon/Amino acid frequencies are utilized for construction of dendrograms. These dendrograms are presented on the side of the heat map. The organism with similar codon/amino acid usage bias tends to cluster together in the dendrogram. The clustering of genomes based on the codon-usage variation reveals two major clades. The one clade encompasses of obligate symbionts like *Frankia* DG and *Frankia* CcI3. The other clade comprises of *Frankia* ACN14a, EAN1pec and Eul1c. However, dendogram based on the amino acid-usage variation divulges a different clustering. Proteomes of Frankia Eul1c and ACN14a have similar amino acid usage are clustered together in one clade while *Frankia* CcI3 and *Frankia* DG belong to another clade. Oddly, proteome of *Frankia* EAN1pec seems to share amino acid usage with that of *Frankia* CcI3 and DG and are therefore clubbed together.

4.4 Homology modeling of nitrogenase iron protein:

4.4.1 Three dimensional model of NifH protein from Bradyrhizobium ORS278:

4.4.1.1 Template selection:

For prediction of the three dimensional structure of a protein, at first a suitable template must be identified and compared with the target protein. The target protein and template protein may be structurally similar if both their sequences and function are alike . Position specific iterative blast (PSI-BLAST) results of the nitrogenase iron

```
CLUSTAL W (1.83) multiple sequence alignment

1NIP            --AMRQCAIYGKGGIGKSTTTQNLVAALAEMGKKVMIVGCDPKADSTRLILHSKAQNTIM
1NIF            MSSLRQIAFYGKGGIGKSTTSQNTLAALAEMGHRILIVGCDPKADSTRLILHAKAQDTIL
                  ::** *:***********:** :*******::::*****************:***:**:

1NIP            EMAAEAGTVEDLELEDVLKAGYGGVKCVESGGPEPGVGCAGRGVITAINFLEEEGAYEDD
1NIF            SLAAAAGSVEDLEIEEVMKVGYRDIRCVESGGPEPGVGCAGRGVITSINFLEENGAYED-
                .:** **:*****:*:*:*.** .::********************:******:*****

1NIP            LDFVFYDVLGDVVCGGFAMPIRENKAQEIYIVCSGEMMAMYAANNISKGIVKYANSGSVR
1NIF            IDYVSYDVLGDVVCGGFAMPIRENKAQEIYIVMSGEMMAMYAANNISKGILKYANSGGVR
                :*:* ***************************** ****************:******.**

1NIP            LGGLICNSRNTDREDELIIALANKLGTQMIHFVPRDNVVQRAEIRRMTVIEYDPKAKQAD
1NIF            LGGLVCNERQTDKELELAEALAKKLGTQLIYFVPRDNIVQHAELRRMTVLEYAPDSVQAG
                ****:**.*:**:* **  ***:*****:*:******:**:**:*****:** *.: **.

1NIP            EYRALARKVVDNK-LLVIPNPITMDELEELLMEFGIMEVEDESIVGKTAEEVA------M
1NIF            HYRNLATKIHGNAGKGIIPTPITMDELEDMLMEHGIMKAVDESQVGKTAADLAATAMSSL
                .** ** *: .*     :**.********::***.***:. *** ***** ::*       :

1NIP            RQCAIYGKGGIGKSTTTQNLVAALAEMGKKVMIVGCDPKADSTRLILHSKAQNTIMEMAA
1NIF            RQIAFYGKGGIGKSTTSQNTLAALAEMGHRILIVGCDPKADSTRLILHAKAQDTILSLAA
                ** *:***********:** :*******::::*****************:***:**:.:**

1NIP            EAGTVEDLELEDVLKAGYGGVKCVESGGPEPGVGCAGRGVITAINFLEEEGAYEDDLDFV
1NIF            AAGSVEDLEIEEVMKVGYRDIRCVESGGPEPGVGCAGRGVITSINFLEENGAYED-IDYV
                  **:*****:*:*:*.** .::********************:******:*****  :*:*

1NIP            FYDVLGDVVCGGFAMPIRENKAQEIYIVCSGEMMAMYAANNISKGIVKYANSGSVRLGGL
1NIF            SYDVLGDVVCGGFAMPIRENKAQEIYIVMSGEMMAMYAANNISKGILKYANSGGVRLGGL
                ***************************** *****************:******.******

1NIP            ICNSRNTDREDELIIALANKLGTQMIHFVPRDNVVQRAEIRRMTVIEYDPKAKQADEYRA
1NIF            VCNERQTDKELELAEALAKKLGTQLIYFVPRDNIVQHAELRRMTVLEYAPDSVQAGHYRN
                :**.*:**:* **  ***:*****:*:******:**:**:*****:** *.: **.**

1NIP            LARKVVDNK-LLVIPNPITMDELEELLMEFGIMEVEDESIVGKTAEEV----
1NIF            LATKIHGNAGKGIIPTPITMDELEDMLMEHGIMKAVDESQVGKTAADLAATA
                ** *: .*     :**.********::***.***:. *** ***** ::
```

Figure 4.17 Alignment of target protein and template protein, with conserved regions shown in boxes. The regions highlighted in red and orange shows the conserved regions around the metal binding sites

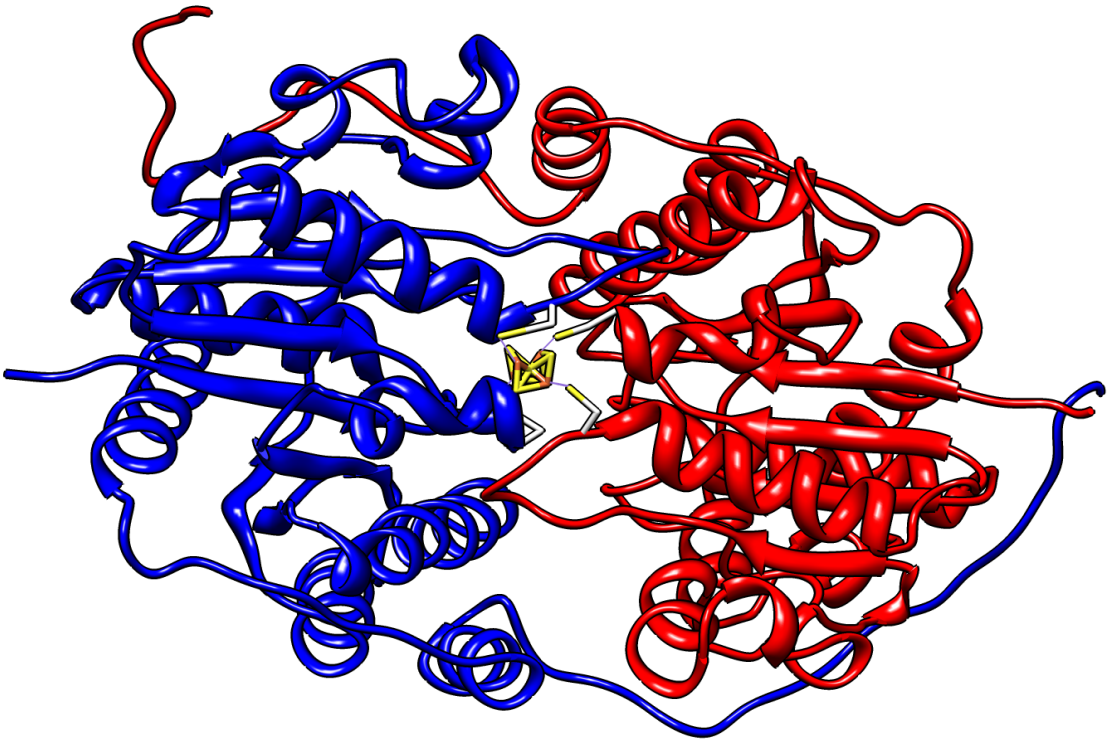


Figure 4.18 3D structure of the nitrogenase iron protein from *Bradyrhizobium* ORS278. The structure comprises of two subunits connected at one surface by the 4Fe-4S cluster

protein (NifH) of *Bradyrhizobium* sp. ORS278 revealed 73% identity with the X ray crystallographic structure of the nitrogenase iron protein from *Azotobacter vinelandii* (PDB ID-1NIP). This protein was chosen as the template for model construction. This template match had an E-value of 3e-151 and X-ray crystallographic resolution of 2.9 Å. The selected protein is a homodimer consisting of 289 residues in each chain. The alignment between the template and target sequence is shown in Figure 4.17.

The alignment reveals major conserved regions in the amino acid residue positions: 10-20, 37-52, 87-106 (with a metal binding site), 125-151(with another metal binding site), 153-169,260-267,305-315,332-347,382-401 (with a metal binding site), 420-446 (with a metal binding site), 448-464, 555-562. The secondary structure prediction servers revealed that the modeled structure has alpha helix which constitutes 256 residues (43.39%), extended strand with 100 residues (16.95%) and the random coil with 234 residues (39.66%) of the protein. All these are interspersed between the protein residues throughout the structure.

4.4.1.2 Features of modeled proteins:

The modeled structure of the nitrogenase iron protein revealed its existence as homodimer with 295 amino acid residues in each chain (Figure 4.18) connected at one surface by the 4Fe-4S cluster. The model has 4428 number of atoms and 4488 bonds. The approx. molecular weight of the modeled protein is 63,575.3 Da and the molecular volume 37,826. The net partial charge on the protein surface was estimated to be -22. Hydrophilic amino acids occupied the surface whereas the hydrophobic amino—acids occupied the core areas. The main functional region is represented by the 4Fe-4S cluster. The cluster is connected by the thiol groups of cysteine residues 99 and 133 in each sub-unit. These residues control conformational reaction of the iron proteins. The structures are chiefly supported by the main-chain atoms of residues 96 to 100 and 130–133, and the hydrophobic side-chains of the residues Ala 100, Gly 101 and Phe 136. These residues are vital for sustaining the suitable cluster setting. The cluster ligands in the structures are located at the amino terminal end of the helices so that amide groups are adjusted in the direction of the cluster

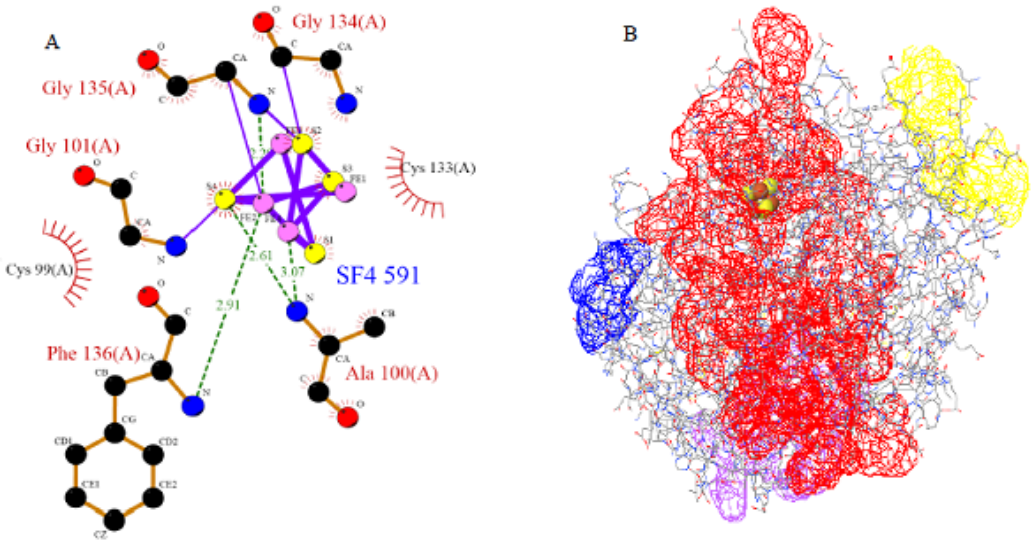


Figure 4.19 (A) Ligand plot of interactions involving the ligands in nitrogenase iron proteins from *Bradyrhizobium* ORS278; (B)Various clefts and cavities present on the surface of the protein. The cleft are highlighted by various color

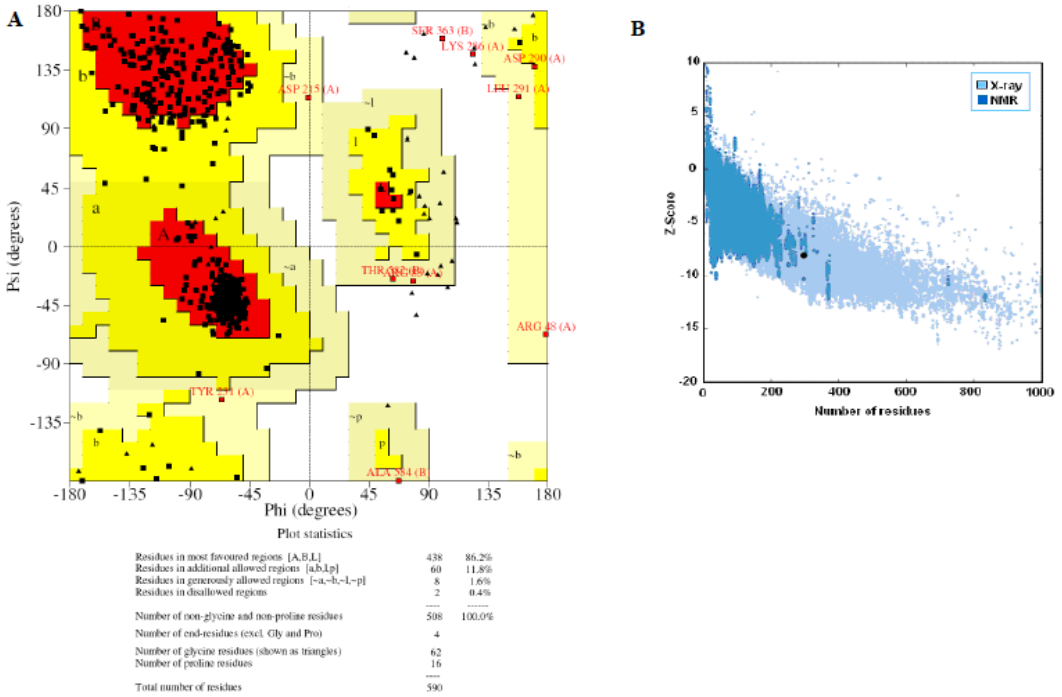


Figure 4.20 (A) Ramachandran plot of the nitrogenase iron protein of *Bradyrhizobium* sp. ORS278; (B) The plot shows z-scores of all protein chains in PDB determined by X-ray crystallography (light blue) or NMR spectroscopy (dark blue) with respect to their length. The plot shows only chains with less than 1000 residues and a z-score -10. The z-scores of NifH protein are highlighted as large dots.

and ligands. Electrostatic interactions are favored. Figure 4.19 A shows a plot of interaction of the ligands. The nucleotide binding section of the nitrogenase iron protein comprise of the residues in the Walker A motif, Switch 1 and Switch 2 regions. The Walker A motif lies between 11-18

residues, Switch I region between 38-44 residues and Switch II region between 124-128 residues in the modeled protein. These regions are associated by polar contacts and indispensable for maintaining the stability of the protein. The switch regions play critical roles in transmitting information concerning the nucleotide state to other effector molecules that bind to these region (Schlessman *et al*.,1998).

CASTp program confirmed the presence 105 pockets for ligands interaction of varying area and volume. Results of ProFunc analysis revealed that the interacting interface of the nitrogenase iron protein from *Bradyrhzobium* sp.ORS278 consist of two chains. There are 36 interface residues in Chain A with an interface area of 2129 angstroms whereas Chain B consists of 42 interface residues with interface area of 2019 angstroms. No disulphide bonds and salt bridges were recognized in the interface region of the protein. The analysis also confirmed the presence of nests in each chain of the protein. Eight such nests have been located in Chain A whereas Chain B consists of 9 nests. These nests are structurally crucial motifs forming a concave depression which can serve as a binding site for an atom or a group of atoms (Watson and Milner-White, 2002). Analysis of the binding sites showed the presence of clefts and cavities in the surface of the proteins (Figure 4.19 B). These clefts and cavities are biologically important. The size of clefts in a protein's surface is one of the key factors determining the interaction of protein with other molecules. It is often seen that in enzymes, the active site is characterized by a particularly large and deep cleft with conserved amino acid residues (Laskowski *et al.*, 2005). Even in the 3D model of the NifH the Fe4S4 cluster was found to be located in a large cleft which further substantiate its functionality.

4.4.1.3 Structure validation:

RMSD calculations established that the nitrogenase iron proteins (NifH) of *Bradyrhizobium* sp.ORS278 had a deviation of 0.52 Å in the Cα residues and 0.55 Å in the backbone residues, from the template protein. This suggests that there isn't much significant deviation between the template protein and the modeled protein. The Ramachandran Plot (Figure 4.20 A) illustrating the backbone conformation of the modeled protein revealed that the number of non

-glycine and non-proline residues was 508. Out of these, 438 (86.2%) residues were in the most favored regions. The refined model was analyzed by different analysis programs for the evaluation of the model quality. The overall quality score calculated by ProSA for our structure is displayed in a plot (Figure 4.20 B) that shows the scores of all experimentally determined protein chains currently available in the Protein Data Bank (PDB) and has a z-score of −8.05. This result indicates that the z-score of our model is very much within the range of scores normally found for proteins of comparable size. ERRAT evaluation of the protein structure revealed a quality factor of 87.34.The normally accepted range of high quality model is <50 (Colovos and Yeates, 1993). In the current case, the ERRAT score is well within the range of a high quality model. The model was further evaluated by VERIFY 3D Structure Evaluation Server and found to be stable and consistent.

It has been previously reported that Alanine-157 is crucial for the nitrogenase iron protein of *Azotobacter vinelandii* to establish the electron-transfer-favored conformation induced by MgATP binding and substitution of this residue with serine (Bursey and Burgess, 1998) leads to complete inactivation of the protein. In the 3D structure of the modeled protein, alanine residue was found to reside in the 158 position. To assess the importance of this residue, single amino substitution was carried out where alanine-158 was replaced by serine *in-silico* and submitted to SDM server. Stability score of this site directed mutagenesis obtained from the SDM server revealed that the mutation bore immense consequences in the functionality of the protein. The results from I-Mutant2.0 analysis reveal that point mutation in the 158 position decreases the thermodynamic stability of the protein. In fact the predicted free energy change values (DDG) were found to be negative upon mutation. However, no significant topographical changes were observed in the modeled protein after mutagenesis. Similarly the replacement of the Fe4S4 cluster binding cysteine residues i.e Cys99 and Cys133 with methionine in structural protein models by *in-silico* methods indicate that there is a decrease in the overall stability of the protein structure due to the mutation. Therefore the *in-silico* site-directed mutagenesis results

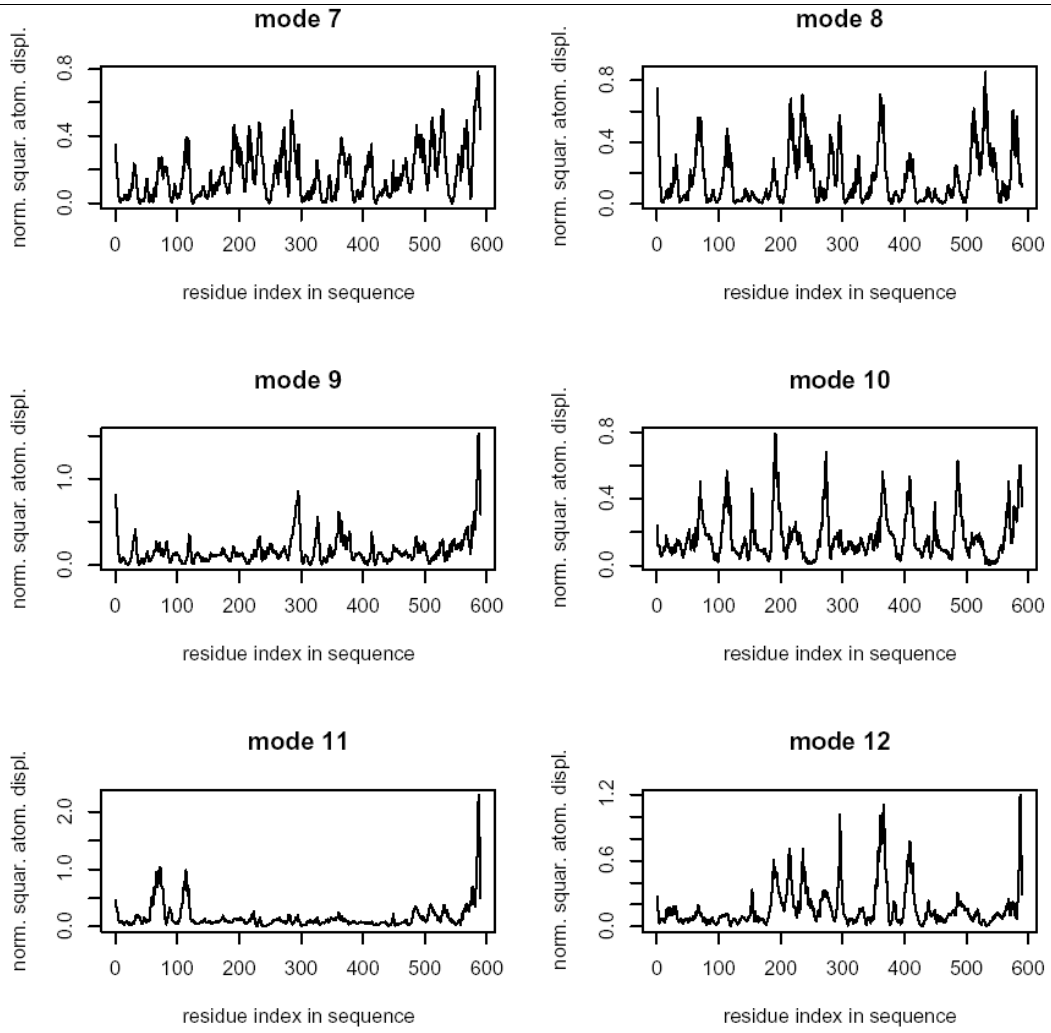


Figure 4.21 Normalized atomic displacement plot calculated for modes 7 to 12 in NifH protein of *Bradyrhizobium* ORS278.

confirm that Ala158, Cys 99 and Cys 133 are functionally important residues. These results also designate that the residues associated with the protein conformations and core functionality is conserved structurally and functionally and any slightest change may have immense influence on the protein functionality.

4.4.1.4 Results of Normal Mode Analysis:

The intrinsic dynamics study of our modeled protein was carried out using the Normal mode analysis. NMA is a powerful technique for studying the vibrational and thermal properties of proteins at the atomic level. The WEBnm@ server offers fast calculation of the 200 lowest frequency modes. The first six modes which are referred to as zero frequency modes correspond to global rotation and translation of the system and are ignored in corresponding analysis (Hollup *et al.*, 2005). Therefore the lowest frequency mode of interest is

mode 7. The deformation energy of the seventh mode of our protein was 1231.61 which were lowest among the other modes. The deformation energy is a measure of the collectivity of the motion associated with the mode. Low deformation energy of the seventh mode indicates that the mode is with large rigid regions which have a good chance of describing domain motions. Normalized atomic displacements analyses were performed for modes 7 to12 (Figure 4.21).The plot thus obtained shows the displacement of each amino acid residue of the modeled protein for a particular mode. The highest correspond to the most displaced regions. A vector field representation of the protein was also generated with WEBnm@ server. The vectors represent the direction and the relative displacements of the different regions of the protein helping us to assess the possible motion of the protein.

4.4.2 *Homology modeling of nitrogenase iron protein from Anabaena variabilis:*

4.4.2.1 Template selection and protein modeling:

The first task in homology modelling technique requires recognition of proper template structure. PSI-BLAST (Altschul *et al.*, 1997) search carried out against database specification of PDB proteins found the crystallographic structure of the nitrogenase iron protein from *Clostridium Pasteurianum* (PDB ID-1CP2) (Schlessman *et al.*, 1998) to be suitable template. The template shared 64% sequence similarity with the nitrogenase iron protein (NifH) of *Anabaena variabilis*. The template protein is a homodimer consisting of 269 residues in each chain. The alignment between the template and target sequence is shown in Figure 4.22.

As seen in the alignment the major conserved region involves residues: 5-20, 40-51, 82-94 (with a metal binding site), 126-132(with another metal binding site), 274-289, 309-320, 351-363 (with a metal binding site), 395-401 (with a metal binding site). It can be seen that region surrounding the metal binding sites are highly conserved. Hierarchical Neural Network (HNN) secondary structure prediction method revealed that the each monomer comprises of alpha helix with 86 residues (31.27%), extended strand with 72 residues (26.18%) and random coil with 117 residues (42.55%) of the protein.

4.4.2.2 Features of the modeled protein and structure validation:

The tertiary structure arrangement of the modeled protein consists of two homodimer subunits that are connected at one surface to the cubane structure of the 4Fe-4S cluster (Figure 4.23 A), which is the core functional region of the protein. Each of the subunit consisted of mixed α-helix/β-sheet

```
CLUSTAL 2.0.9 multiple sequence alignment

1CP2            MRQVAIYGKGGIGKSTTTQNLTSGLHAMGKTIMVVGCDPKADSTRLLLGGLAQKSVLDTL
1ANA            MRKIAIYGKGGIGKSTTTQNMVAGLVEIERKVMVVGCDSKADSTRLLLGGLHPKIVLDTL
                **:.**************.::**  : :.:******.**********   * *****

1CP2            REEGEDVELDSILKEGYGGIRCVESGGPEPGVGCAGRGIITSINMLEQLGAYTD--DLDY
1ANA            RQESDDLHLEDLCKEGWGKTRCVESGGPEPGVGCTGRGVITSIGLLEQLGAYDEKLQLDY
                *:*.:*:.*:.: ***:*   **********.:***:****.:******* :  :***

1CP2            VFYDVLGDVVCGGFAMPIREGKAQEIYIVASGEMMALYAANNISKGIQKYAKSGGVRLGG
1ANA            TFYDGLGSVVCGGFVMPMRERKAQEVYIVTSGEIMAMYTTNNICRGIHKYTVIGGVRLGG
                .*** **.******.**:** ****:***:***:**:*::***.:**:**:  *******

1CP2            IICNSRKVANEYELLDAFAKELGSQLIHFVPRSPMVTKAEINKQTVIEYDPTCEQAEEYR
1ANA            LICNSRNIEQEDDLVKAFAEELGTQMIAFIPRDKIVQQAEFYSQTVIEYAPHCEQAQHYR
                :******:: :* :*:.***:***:*:* *:**. :* :**:  .****** * ****:.**

1CP2            ELARKVDANELFVIPKPMTQERLEEILMQYG----MRQVAIYGKGGIGKSTTTQNLTSGL
1ANA            NLAKAIDQNTNFVIPKPLSDEQLEELLVNSGVLSVMRKIAIYGKGGIGKSTTTQNMVAGL
                :**: :* *  ******:::*:***:*:: *    **::**********:.::**

1CP2            HAMGKTIMVVGCDPKADSTRLLLGGLAQKSVLDTLREEGEDVELDSILKEGYGGIRCVES
1ANA            VEIERKVMVVGCDSKADSTRLLLGGLHPKIVLDTLRQESDDLHLEDLCKEGWGKTFCVES
                  : :.:******.**********  * ******:*.:*:.*:.: ***:*   ****

1CP2            GGPEPGVGCAGRGIITSINMLEQLGAYTD--DLDYVFYDVLGDVVCGGFAMPIREGKAQE
1ANA            GGPEPGVGCTGRGVITSIGLLEQLGAYDEKLQLDYTFYDGLGSVVCGGFVMPMRERKAQE
                *********:***:****.:******* :  :***.*** **.******.**:** ****

1CP2            IYIVASGEMMALYAANNISKGIQKYAKSGGVRLGGIICNSRKVANEYELLDAFAKELGSQ
1ANA            VYIVTSGEIMAMYTTNNICRGIHKYTVIGGVRLGGLICNSRNIEQEDDLVKAFAEELGTQ
                :***:***:**:*::***.:**:**:  *******:*****:: :* :*:.***:***:*

1CP2            LIHFVPRSPMVTKAEINKQTVIEYDPTCEQAEEYRELARKVDANELFVIPKPMTQERLEE
1ANA            MIAFIPRDKIVQQAEFYSQTVIEYAPHCEQAQHYRNLAKAIDQNTNFVIPKPLSDEQLEE
                :* *:**. :* :**: .****** * ****:.**:**: :* *  *******:::*:***

1CP2            ILMQYG----
1ANA            LLVNSGVLSV
                :*:: *
```

Figure 4.22 Alignment of target protein and template protein, with conserved regions shown in boxes.

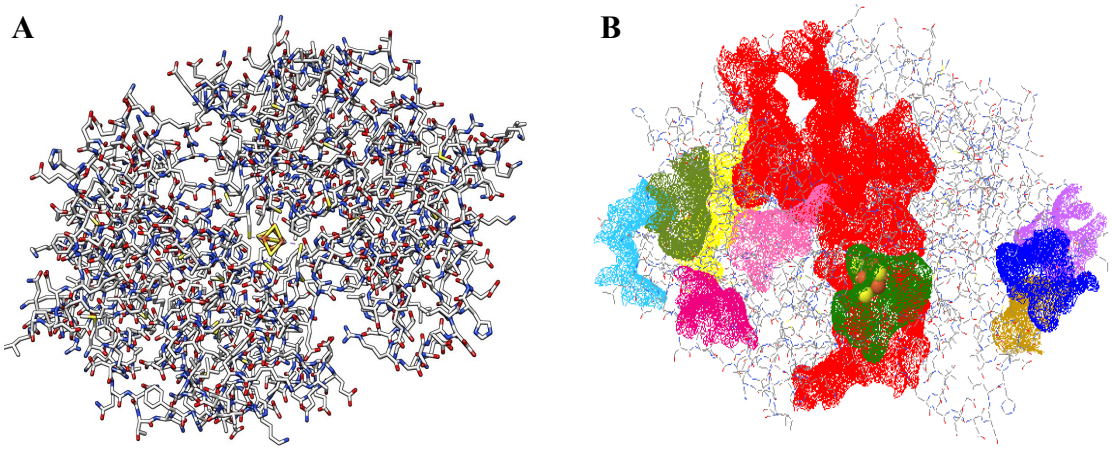
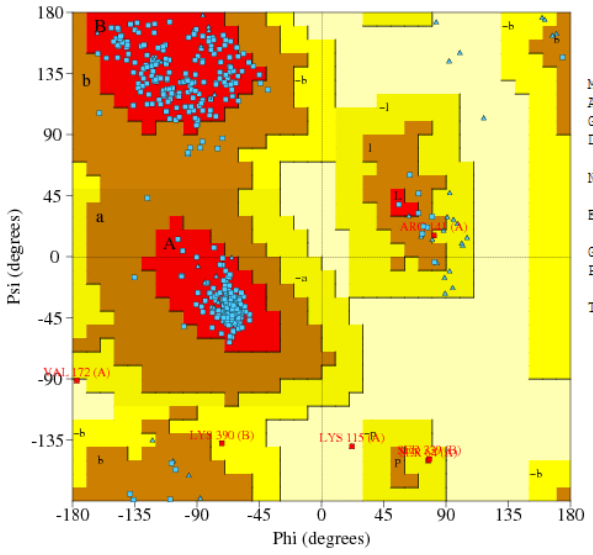


Figure 4.23 (A)3D structure of the nitrogenase iron protein from *Anabaena variabilis*; (B) Predicted cavities on the protein surface

polypeptide fold. Ligating cysteine residues are Cys 94 and Cys 131 which connects the metal cubane cluster with main body. The model has 4236 number of atoms. The exact molecular weight of the target protein is 60,722 Da and the molecular volume 35,994. The net partial charge on the protein surface was estimated to be -23. Profunc analysis revealed 10 different clefts and cavities with varied volume in the protein structure (Figure 4.23 B). The largest cavity comprises of 85 residues with volume 12621.23 and a cavity also exists around the Fes4S4 metal cluster. These cavities or clefts are the sites where small molecules bind and thus they constitute the functional sites of the protein. Functional analysis also revealed presence of nests on the protein surface. 7 nest are located in chain A whereas 8 nest are located in chain B. These nests are structurally important motifs found in functionally important regions of the protein structures. These nests appear like concave depressions serving as binding sites for atoms. Like



		No. of residues	%-tage
Most favoured regions	[A,B,L]	435	92.6%
Additional allowed regions	[a,b,l,p]	29	6.2%
Generously allowed regions	[~a,~b,~l,~p]	4	0.9%
Disallowed regions	[XX]	2	0.4%*
Non-glycine and non-proline residues		470	100.0%
End-residues (excl. Gly and Pro)		5	
Glycine residues		60	
Proline residues		16	
Total number of residues		551	

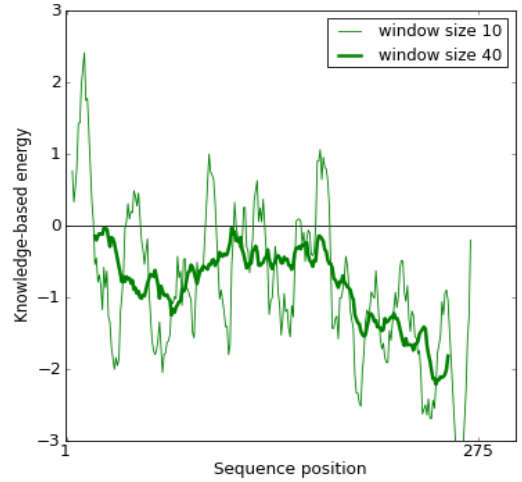


Figure 4.24 (A) Ramachandran plot of the nitrogenase iron protein of *Anabaena variabilis*; (B) Energy plot of the protein. Residue energies are averaged over a sliding window are plotted as a function of the central residue in the window

the NifH protein from *Bradyrhizobium,* this protein also has a Walker A motif (residues 8-15) and two switch regions, Switch I (residues 35-41) and Switch II (residues 122-126) that interact with the γ- phosphate group of the bound nucleoside triphosphate. No potential helix turn- helix DNA binding motifs were identified in the protein. CASTp program confirmed the presence 83 pockets for ligands interaction of varying area. Pockets on the protein surface crafts the inherent physio-chemical features indispensable for NifH protein functioning.

The quality of the refined model assessed by different programs like ProSA, Verify 3D and ERRAT Score suggested that the predicted protein model was quite reliable and consistent. The overall quality score determined by ProSA for the structure showed that its score were consistent with the scores of experimentally determined protein chains available in the Protein Data Bank (PDB). The NifH protein of *Anabaena variabilis* had a z score of -7.08. This result pointed out that the z-score of our model was within the range of scores normally found for proteins of comparable size. Interestingly the energy distribution plot in Figure 4.24 B (using a window size of 40 as default) of the modeled protein is below the zero base line. This result of the energy plot suggests that the predicted protein model is quite consistent. ERRAT analysis revealed that the overall quality factor of protein structure was 91.38. The Ramachandran Plot (Figure 4.24 A) illustrating the backbone conformation of the modeled protein revealed that the number of non-glycine and non-proline residues was 479. Out of these, 435 (92.6%) residues were in the most favored regions. Moreover, the percentages of residues in the additional allowed region and generously allowed region are 6.2% and 0.9 % respectively. However, 0.4% of residues remain on the disallowed region.

4.5 Molecular dynamics simulation:

4.5.1 MD simulation of dimer of NifH protein from Bradyrhizobium ORS278:

The molecular dynamics simulation of

Table 4.7 Partial charges of the iron–sulphur clusters as obtained from *ab initio* calculations

Atom	Charge
Fe1	0.758303
Fe2	0.755864
Fe3	0.776678
Fe4	0.791489
S1	-0.259653
S2	-0.264117
S3	-0.284825
S4	-0.273739

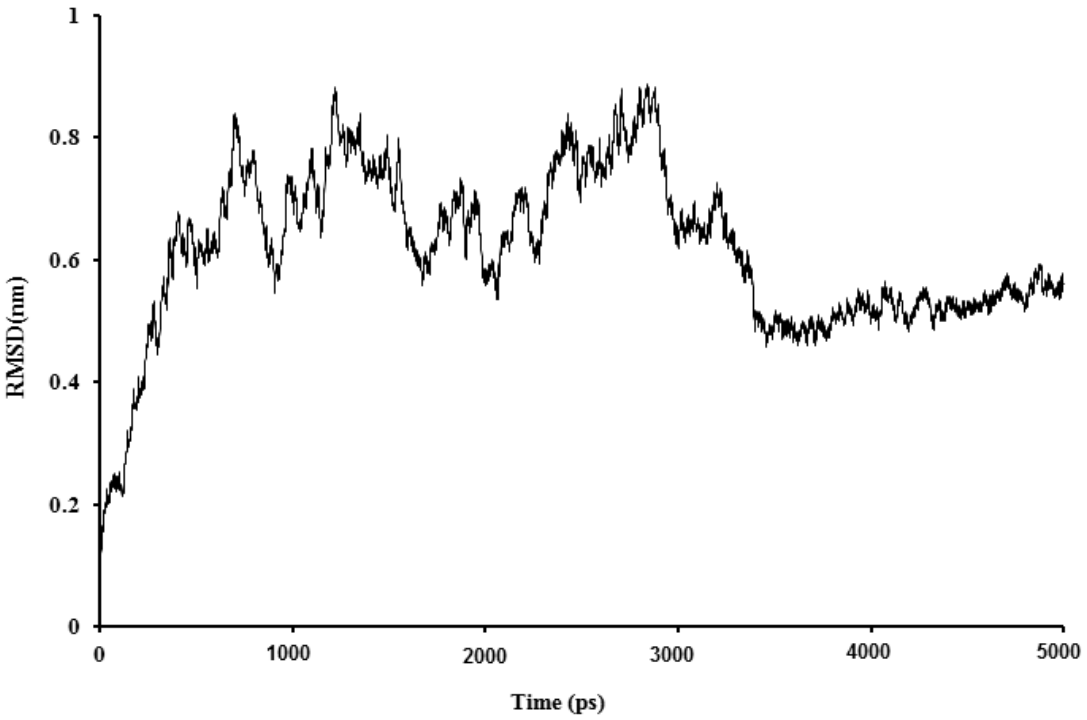


Figure 4.25 The time evolution of RMSD during 5000ps dynamics simulation time

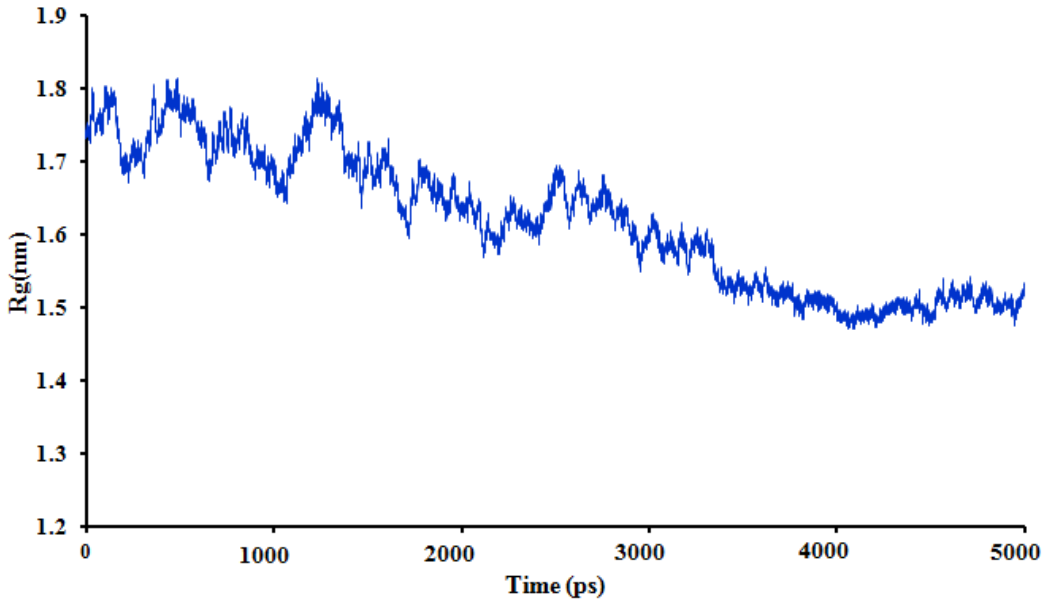


Figure 4.26 Time evolution of changes in radius of gyration of the protein in aqueous medium

homodimeric NifH protein from *Bradyrhizobium* ORS278 with attached iron-sulfur cluster (Fe4S4) was carried for 5 ns to explore the molecular conformational changes as a function of time. Before the actual MD run, the parameterization of the metal centre was carried out. The partial charges of the iron-sulfur cluster determined by quantum chemical calculations are

provided in Table 4.7

The overall stability of the protein during the simulation was monitored using several parameters like Root Mean Square Deviation (RMSD), Radius of gyration (Rg), Root Mean Square Fluctuation (RMSF), etc. Time-dependent Cα rms deviation (RMSD) was used to provide a picture of the global drift of the homology model from the initial protein structure during the simulation period (Figure 4.25). It is evident from figure that RMSD increased gradually in the early stages but there are quite a bit of fluctuation in the period 1000-3000 ps and then there is a leveling off in deviation from 3500 ps onwards. The initial rise of the RMSD over is possibly attributable to the relaxation of motion of the protein in the solution.

The time evolution of Radius of gyration (Rg) is presented in the Figure 4.26. The Radius of gyration gives us an idea about the compactness of the structure. The figure clearly shows that there is quite a variation in the Radius of gyration during the simulation particularly during 1500 to 2500 ps time frame. This further proves that a conformational change is taking place during the simulation period. From 3000 ps onwards no major deviation in Rg is noticed which indicates that the protein is not much flexible.

The root mean square fluctuation is a measure of flexibility of each residue over the trajectory. Figure 4.27 demonstrates the RMSF of Cα atoms plotted against the residue number over the time period 1500-3000 ps. This time interval was chosen because a probable conformational change was noticed in this period in the previous analyses. The figure depicted that RMSFs for each chain adopts largest values at the C terminus where RMSF greater than 0.5 nm. Such large fluctuation value observed for the C terminus is attributable to the presence of unstructured free end, particularly in chain B. In the central region of the protein, consisting of a set parallel β-sheets flanked by α-helices, no significant deviation in RMSF is noticed between the two chains. This stable behavior of the central region is attributable to the network of hydrogen bonds stabilizing these secondary structures. However, a major fluctuation is noticed in the residues 170-180 in chain A and residues 30-40 in chain B. This indicates that this region of the protein shows most mobility in this period. Some of these residues are part of the Switch region

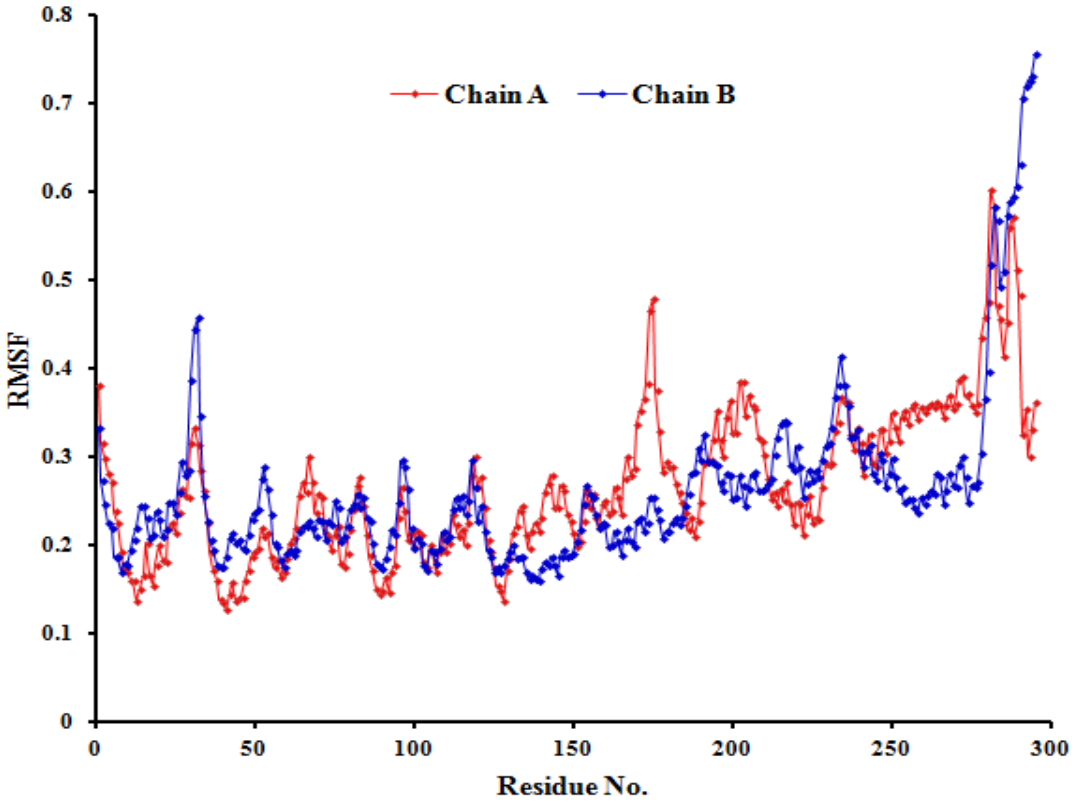


Figure 4.27 Fluctuation in the Cα atoms of the residues of the protein in aqueous medium

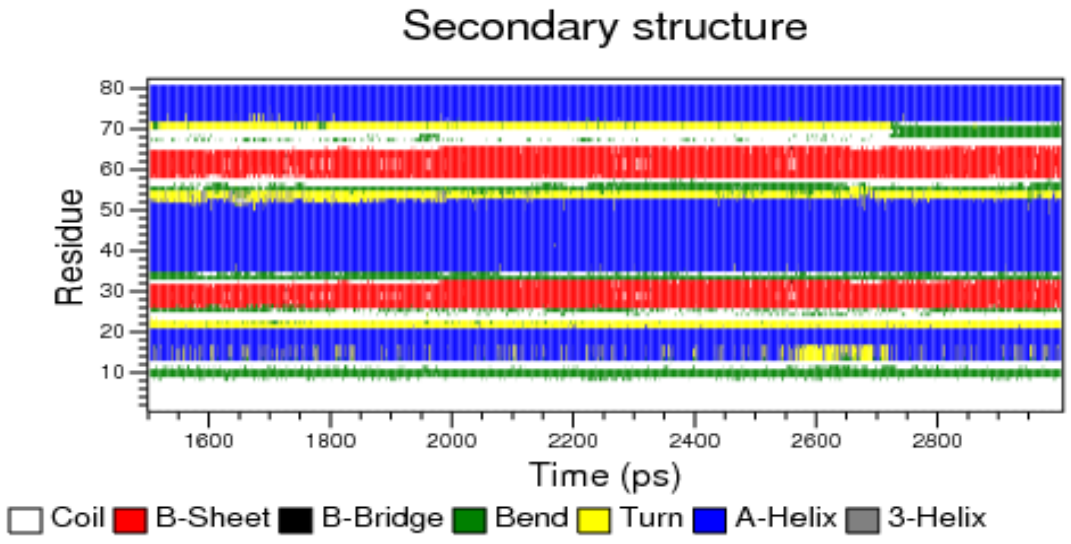


Figure 4.28 Secondary structure evolution of protein during time frame 1500-3000 ps

of the protein. The switch regions are known to undergo conformational changes upon their interaction with γ-phosphate group of the bound nucleoside triphosphate (Lahiri *et al.*, 2008). Thus, this region adopts a structural flexibility. Additional information on the structural flexibility is obtained through the analysis of time -dependent secondary structure

fluctuations. Analysis of the Figure 4.28 reveals very high stability of the motifs with the well defined secondary structure. Thus, β sheets and α -helices observed within the NifH structure are very stable during the whole simulation period, whereas beta turns become bends, and vice versa. The lowest time-dependent stability and lowest structure conservation is observed for the C terminal end. This observation is in agreement with the results of RMSF analysis. There was no significant influence on the total energy of the system. During the time course of the simulation, the total potential energy did not exceed -9.6 x 10^5 KJ mol^{-1}, indicating that protein was energetically stable during simulation.

4.5.2 *MD simulation of monomer of NifH protein from Anabaena variabilis:*

The molecular dynamics simulation of monomer NifH protein from *Anabaena variabilis* with attached iron-sulfur cluster (Fe4S4) was carried for 10 ns to unveil the motional properties of the monomer protein. The charges for the metal cluster determined by the quantum chemical calculation are provided in Table 4.8

Table 4.8 Partial Charges calculated for Fe4S4 cluster in monomer of NifH of *Anabeana*

Atom	Charge
Fe1	0.721409
Fe2	0.727454
Fe3	0.725753
Fe4	0.734837
S1	-0.228108
S2	-0.230229
S3	-0.234819
S4	-0.216298

Several parameters such as RMSD, Radius of gyration (Rg), RMSF, etc, were considered to assess the structural stability of the protein during the course of simulation. Time-dependent Cα rms deviation (RMSD) has been used to provide a picture of the global drift of the homology model during the simulation period. Thus, during the simulation, the RMSD drift of Cα atoms from the initial protein structure was determined. Figure 4.29 demonstrated that the drift observed for the monomeric structure. The curve reaches a plateau after 4100 ps of simulation. The initial rise of the Cα RMSD over the first 1000 ps is possibly attributable to the relaxation motion of the protein in the solution.

The RMSF distribution observed (Figure 4.30) for monomeric structure has much sharper peaks indicating the residues are more flexible than those observed for the previously described dimeric NifH structure. The more stable behaviour of the homodimeric structure can be explained by the

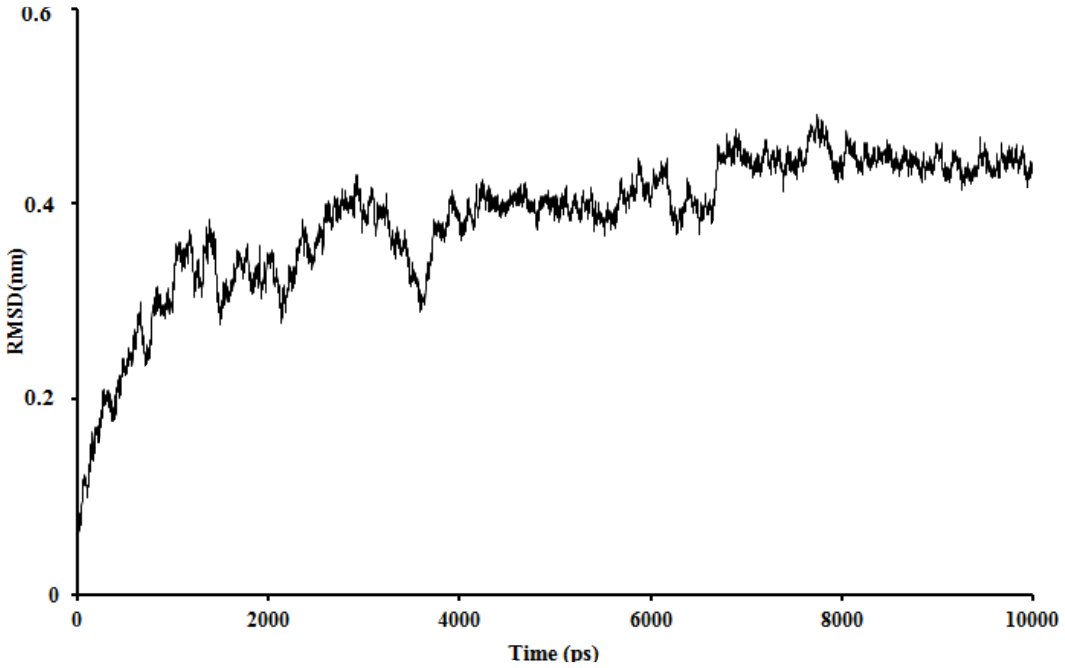


Figure 4.29 Drift of protein structure from the initial model. The RMSD of all Cα atoms from the starting structure is shown as a function of time

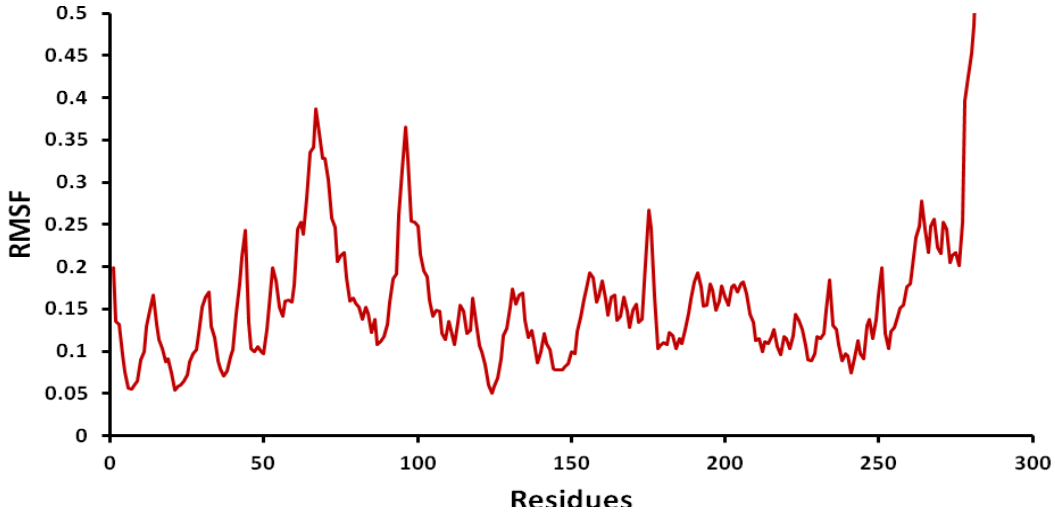


Figure 4.30 RMS fluctuation in the Cα atoms of the residues of the monomeric protein in aqueous medium

nature of the interactions between the units of the homodimer and by the wrapping of the C-terminal ends of each monomer around one another. Additional information on the structural flexibility is offered by the analysis of time-dependent secondary structure fluctuations of the whole monomer unit during the 4000-6000 ps time frame. The time frame was selected to examine the changes in the secondary structure because the protein was structurally quite stable in this period. Analysis of Figure 4.31 reveals

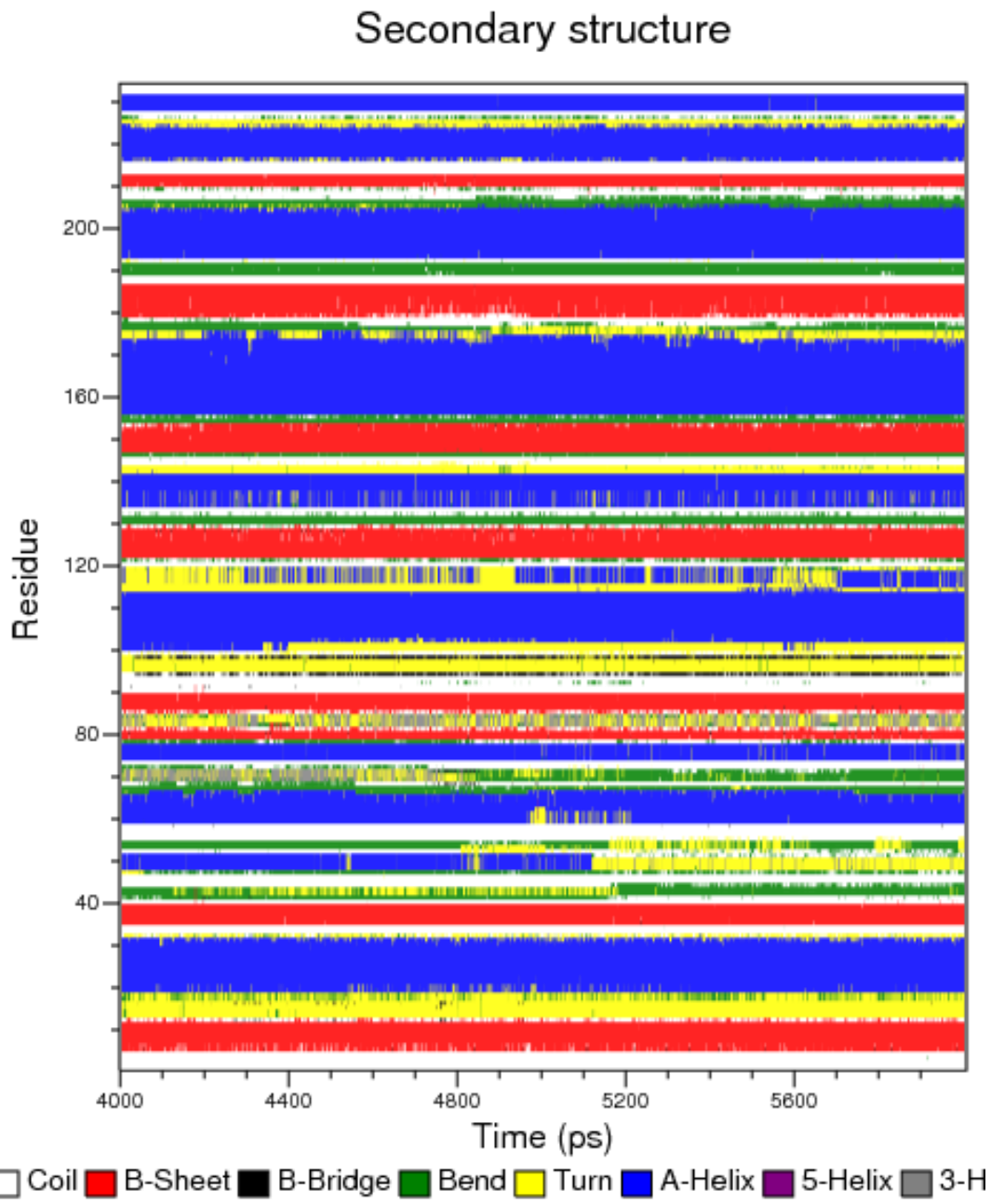


Figure 4.31 Secondary structure evolution of the monomer protein during the time frame 4000-6000 ps

very high stability of the motifs with the well defined secondary structure. Thus, α-sheets and β-helices observed within structure are very stable during the whole simulation period, whereas turns become bends, and vice versa. No major variation in the radius of gyration was noticed during this time period of simulation and the total energy of the subunit also remain constant during most of the simulation period, indicating that the 3D of the NifH monomer is quite stable.

4.6 Molecular docking experiment:

Reversible ADP-ribosylation of dinitrogenase reductase at Arginine

residue is a important mode of post translational regulation of nitrogenase enzyme. These events are probed through *in-silico* docking experiment. Initial sequence analysis of NifH (Fe protein) of various nitrogen fixing organisms revealed that the region around the Fe4S4 cluster is highly conserved and in most of the cases arginine residue is placed near the metal cluster tucked between two glycine residues. Further a close inspection of the structural details of the NifH protein model of the *Bradyrhizobium* ORS278 reveals that probable site of the post translational modification (Arg102) falls within a region of the protein predicted to possess β-sheet or α-helical structure and the arginine residue is only three amino acids removed from one of the cysteine residues which is involved in the binding of the iron-sulfur center to the protein

In-silico docking of the substrates (ADP-Ribose) against whole structures of NifH protein of *Bradyrhizobium* sp. ORS278 produced 3 clusters of conformers using RMSD-tolerance of 2.0A^{o} out of 100 docking runs. The conformation of the #1 ranked cluster was favored in that structure and repeated 83 times out of 100 runs. The substrate binds the Fe protein in their cavities with the lowest docking energies of -5.10 Kcal/mol (Figure 4.32). The lowest binding energy is correlated with highest binding affinity in these substrate-protein interactions. Other than Arg 102, the amino acid

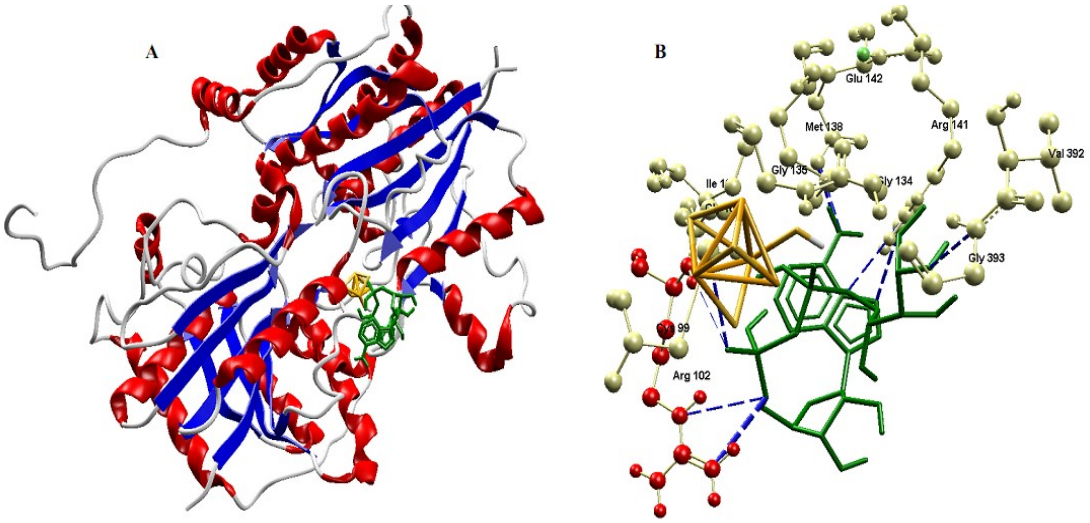


Figure 4.32 (A) The ADP-ribose docked into the three dimensional structure of NifH protein of *Bradyrhizobium* ORS278 ; (B) The docked conformation of the substare ADP-ribose (green). All the amino acid residues which involved in molecular interaction are shown in ball and stick drawing. The arginine residue 102 is shown in Red. All the hydrogen bonds are shown in blue dashed lines

residues found to be involved in molecular interaction with substrate include Cys 99, Arg 141, Met 138 and Gly 393. The possible hydrogen bond of the substrate ADP-Ribose with the amino acid is shown in the figure with blue dashed lines. ADP-Rib is linked to the arginine residue 102 through the guanidino group. Proximity of the docked substrate to the Fe4S4 active center leads to the speculation that this kind of modification may lead to blocking of the binding site for the MoFe protein or form a conformational change that inhibits the protein-protein interaction. As a result, the whole nitrogenase complex may be inactivated. Therefore, ADP-ribosylation of Arg102 may serve as important regulatory system in nitrogenase activity of *Bradyrhizobium* ORS278.

4.7 Phylogenetic analysis of *nif*HDK using alternative approaches:

4.7.1 Sequence and Codon-triplet based phylogeny:

Two different approaches were used to outline the evolutionary relationship of *nif*H, D and K genes. The first comprises of the classical amino-acid sequence based phylogeny and other is the novel codon triplet based phylogeny. Figure 4.33 and Supplementary Figure SF4.1-4.5 represents the dendrograms constructed for *nif* genes using above mentioned ways. The distance matrices generated in the triplet based method are given in Supplementary Table ST8-10.

The pattern obtained in the phylogram for *nif*H (Figure 4.33) based on leading eigen values seem interesting at very first glace. The placement of *Frankia* ACN14a away from the other actinobacteria and *Synechococcus* sp. JA-3-3Ab being isolated from rest of cyanobacterial strains is apparently quite interesting. Members of various classes of Proteobacteria (alpha, beta, gamma and delta) are clustered together in the triplet based phylogenetic tree. However, in amino acid sequence based tree, Delta proteobacteria like *Geobacter, Desulfuromona*s etc. and Gamma proteobacteria members like *Azotobacteria, Klebsiella* etc. remains totally separated from the alpha proteobacteria clade. Another fascinating thing is the placement of *Anabeana variabilis*. In case of phylogenetic tree based on amino acid sequences (Supplementary Figure SF 4.1), it was noticed that NifH of *Anabaena variabilis* shows affinity with firmicutes like *Alkaliphilus* and

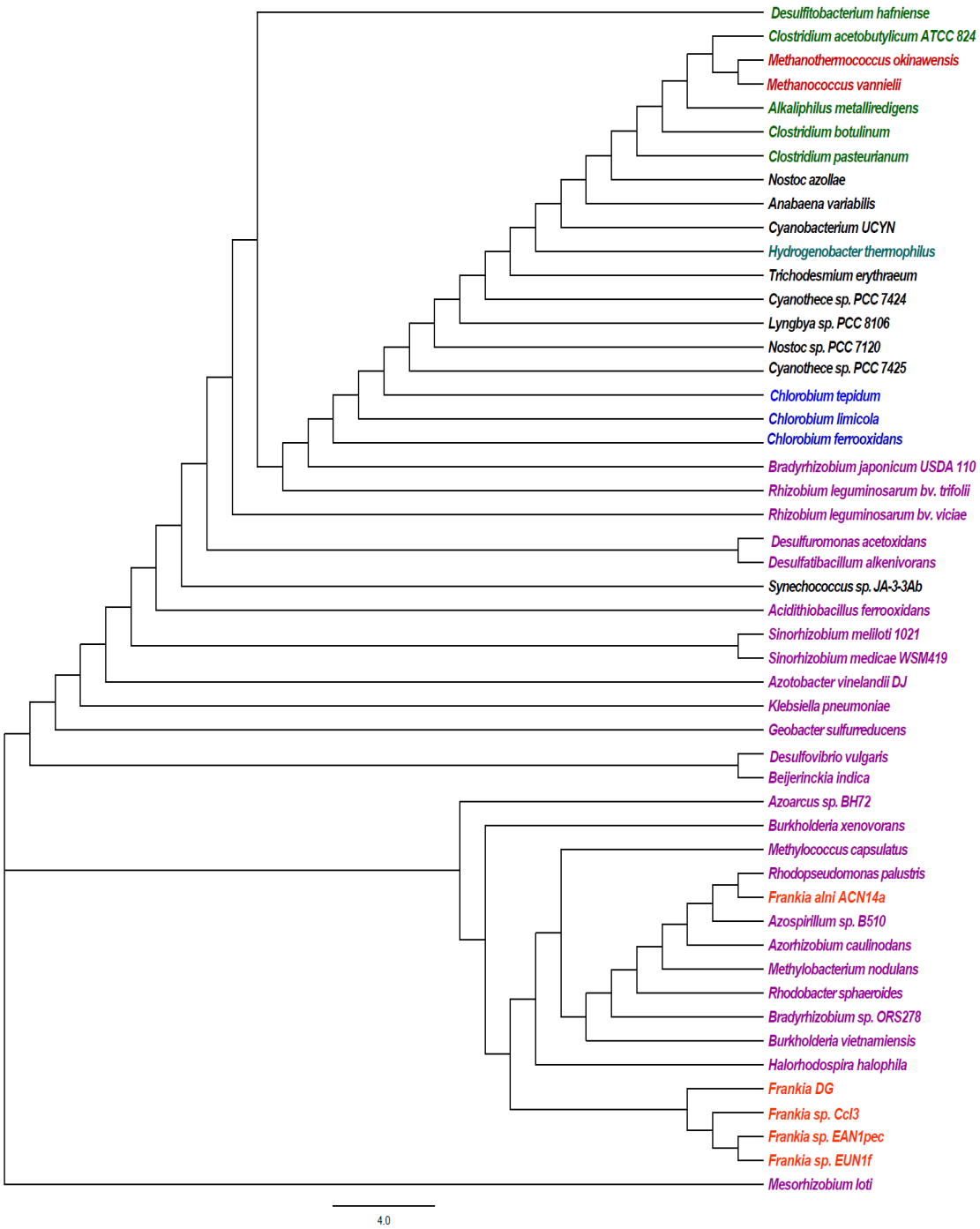


Figure 4.33 Phylogram of *nif*H gene based on codon triplet method. Coloured fonts are used to indicates different classes of diazotrophs. Purple is used for proteobacterial strains; Black is used for cyanobacterial strains; Blue for Green-sulphur;Orange for Actinobacteria; Green for Firmicutes; Red for methanogenic and Grey for Aquificae.

Clostridium whereas in triplet based phylogenetic tree, it is placed amongst *Nostoc azollae* and *Cyanobacterium* UCYN-A. In both type of *nif*H phylogenetic tree, *Hydrogenobacter thermophilus* lie close to the cyanobacteria. The clustering of *Frankia*, an actinorhizal symbiont, with cyanobacterial phyla in the NifH sequence based phylogeny is

conspicuous and has been discussed previously (Normand *et al.*, 1992; Hirsch *et al.*, 1995).

In amino-acid based dendrogram of *nif*D (Supplementary Figure SF4.3), a clear distinction is noticed between the alpha proteobacteria and rest of the proteobacteria members. Most of the cyanobacteria, actinobacteria and firmicutes have clustered together in this tree. However in the tree based on condensed matrix method (Supplementary Figure SF4.2), mottled distribution of *nif*D among cyanobacteria was noticed which reflects the dilemma of these organisms in carrying out both nitrogen fixation and oxygenic photosynthesis at the same time. Diazotrophic cyanobacteria have evolved several mechanisms for segregating nitrogen fixation from O_2 (Berman-Frank *et al.*, 2003) which is reflected in its phylogeny. The placement of proteobacteria members in this tree echoes their polyphyletic origin. A distinct separation between proteobacterial classes is not obvious; for example, *K. pneumoniae* (ɣ) and *M. loti* (α) cluster together. The green-sulphur and methanogeneic bacteria remain in close by but the clostridial members lie in different clades in this tree. Similarly in the codon-triplet based phylogeny of *nif*K (Supplementary Figure SF4.4), different classes of proteobacteria were placed together. Some members of cyanobacteria were also placed along with proteobacteria suggesting they are not monophyletic in origin. The sporadic pattern of distribution of the *nif*HDK genes in various diazotroph as observed in the condensed matrix based phylogram supports the observation made by previous workers (Normand and Bousquet, 1989; Kechris *et al.*, 2006) that the core nitrogen fixation genes are horizontally transferred. Another possible reason for the mosaic pattern exhibited in the placement of diazotroph is the sharing of common ancestry followed by gene-duplication events as suggested by Fani *et al*., (2000). Results obtained in the codon-triplet based phylogeny are better placed that those obtained in the sequence based phylogeny in explaining complex events leading to evolutionary process of nitrogen fixation.

4.7.2 Structure based evolutionary relationships:

NifH protein has diverse range of homologs that are not known to be directly involved with fixing nitrogen.

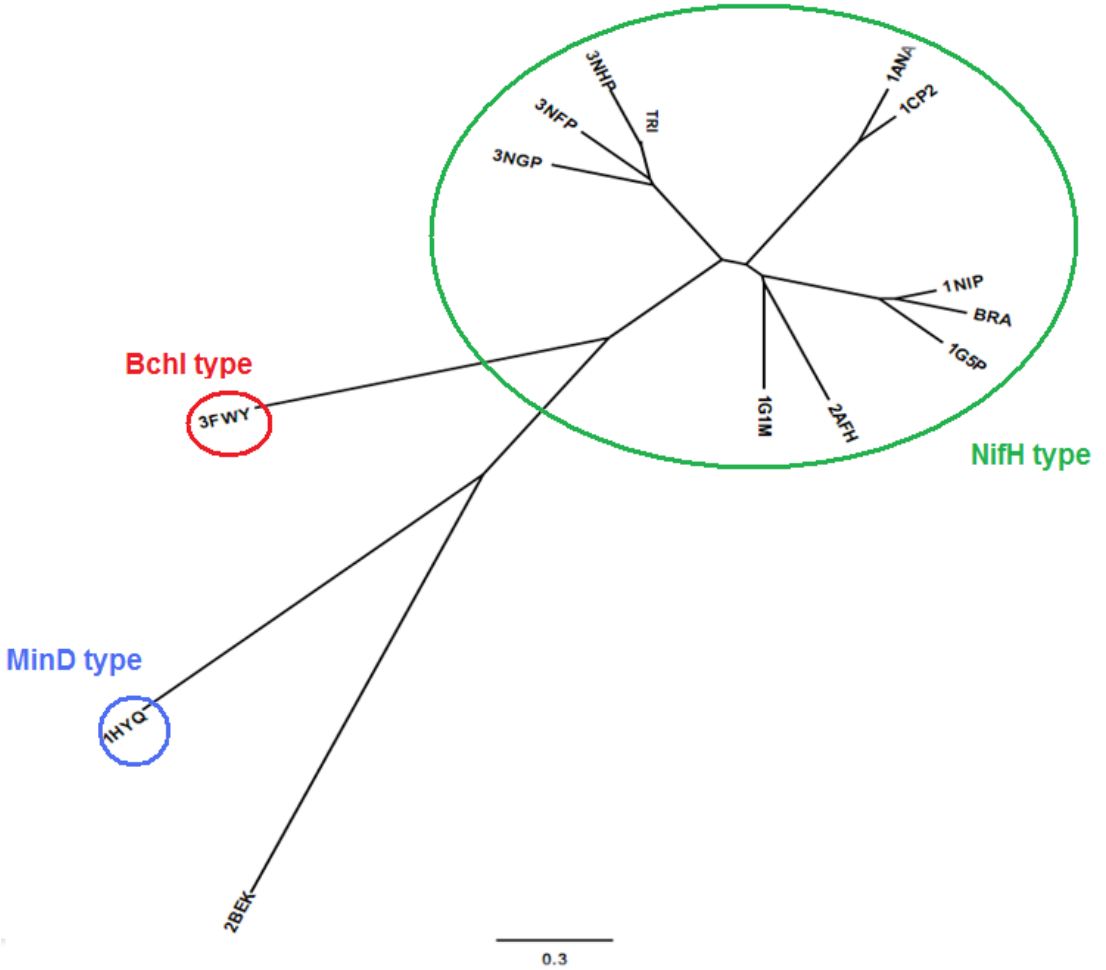


Figure 4.34 Structure-based phylogenetic tree of NifH/BchL family inferred from the root mean square distances

These include metallo proteins like BChL, MinD-1 and other chromosomal segregation proteins. Sequence similarities between Nitrogenase iron protein (NifH) and other metallo proteins has led to the speculation of whether all these proteins have diverged from a single common ancestor or from one of the proteins amongst this group. In order to understand the evolutionary relationship between these proteins and trace their ancestry, a phylogram based on their three dimensional structures was generated. The evolutionary tree based on the structure (Figure 4.34) suggests that NifH protein and ChlL (3FWY) share maximum similarities and have probably diverged most recently compared to others. It also suggests that both NifH and ChlL could have evolved from a MinD like ancestor (1HYQ), a bacterial cell division regulator protein which evolved parallelly with Soj chromosome segregation protein (2BEK). It was also found that NifH proteins *Frankia* strains are relatively

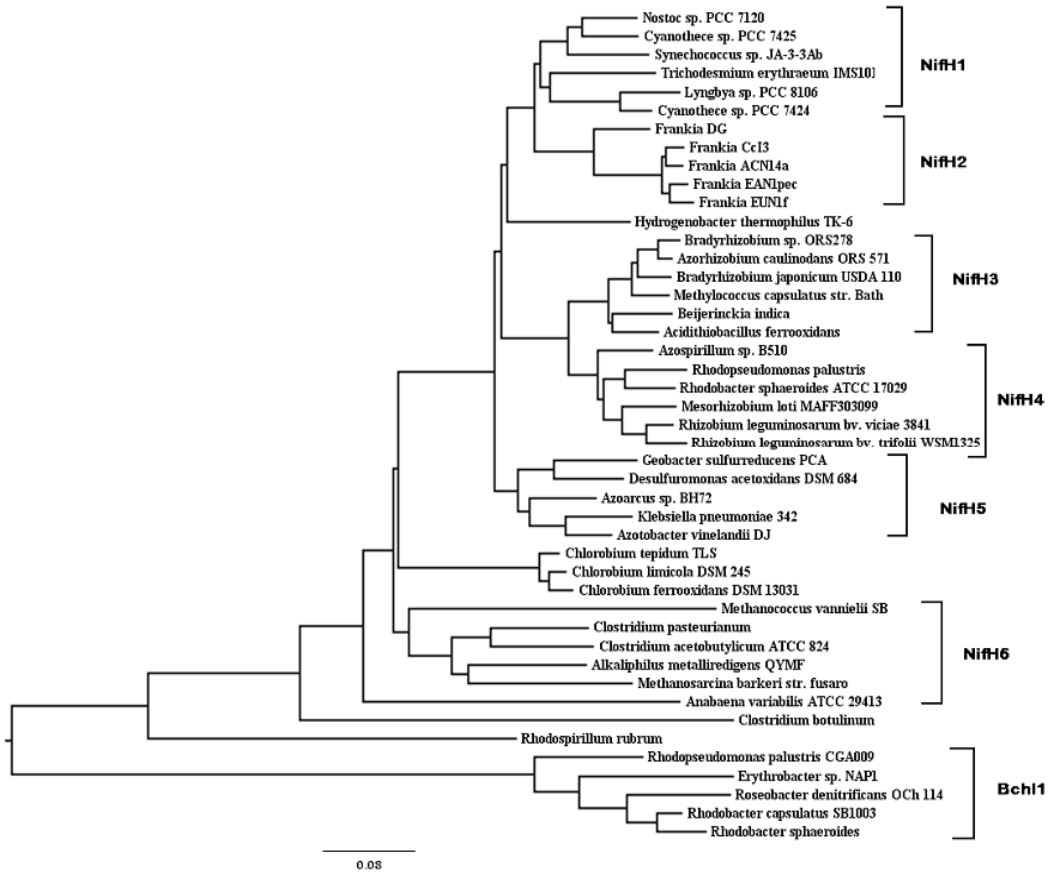


Figure 4.35 Neighbour joining-type phylogenetic tree based on amino acid sequence of the members of NifH/BchL protein family

close to that of the cyanobacteria *Trichodesmium erythraeum* while the structure of NifH of another cyanobacterium, *Anabaena variabilis* is apparently quit apart and shows proximity with that of *Clostridium pasterium*. The NifH protein of *Bradyrhizobium* is structurally quite closer to that of *Azotobacter vinelandii* (1NIP and 1G5P).

4.8 Functional divergence in NifH/ BChl protein family:

4.8.1 Analysis of functional divergence:

A neighbor-joining type phylogenetic tree based on the amino acid sequences was produced to assess the evolution of various members of the NifH/BchL family (Figure 4.35). Here the NifH sequences form a clearly defined clade, but there are also a large number of more distant relatives, these include proteins involved in the synthesis of photosynthetic pigments, namely protochlorophyllide reductase (BchL or ChlL) and chlorin reductase (BchX). Based on clustering, the phylogenetic tree has been further divided into six major sub-clusters of NifH and one of BchL sequences. NifH1 subcluster

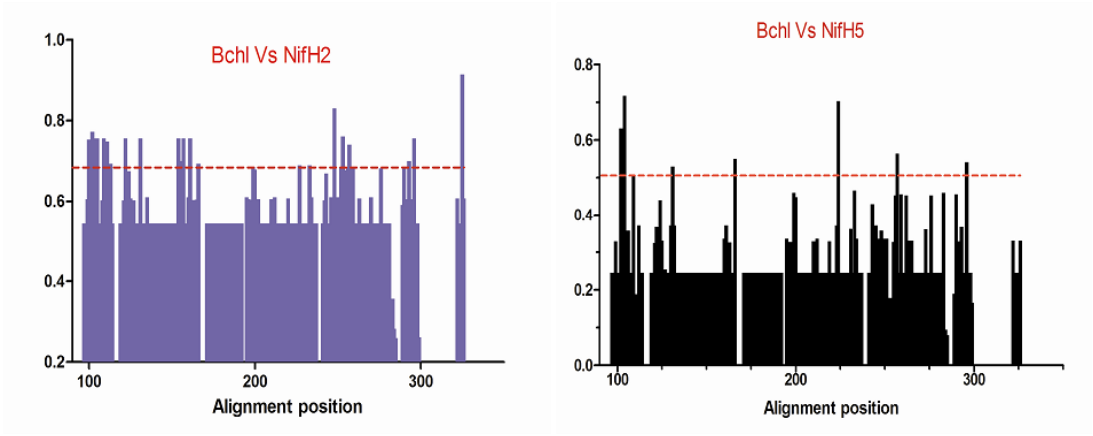
Table 4.9 Estimates of the coefficient of functional divergence (θ) in various cluster pairs

Comparison among the subclusters	θ ML	SE θ	αML	LRT θ
Bch1/NifH1	0.4408	0.1221	0.4418	13.0437
Bch1/NifH2	0.5712	0.1741	0.4188	10.7592
Bch1/NifH3	0.5632	0.1292	0.3242	18.9933
Bch1/NifH4	0.4656	0.1323	0.4045	12.3788
Bch1/NifH5	0.2768	0.1290	0.4351	4.6077
Bch1/NifH6	0.5656	0.1130	0.5985	25.0620
NifH1/NifH2	0.0960	0.1526	0.2887	0.3958
NifH1/NifH3	0.2216	0.0921	0.2327	5.7884
NifH1/NifH4	0.2056	0.1235	0.3048	2.7694
NifH1/NifH5	0.0744	0.0893	0.3089	0.6942
NifH1/NifH6	0.1592	0.0750	0.4775	4.5002
NifH2/NifH3	0.4917	0.1424	0.1576	11.9189
NifH2/NifH4	0.4944	0.1423	0.2255	12.0680
NifH2/NifH5	0.0952	0.1242	0.2588	0.5871
NifH2/NifH6	0.2032	0.1170	0.4952	3.0182
NifH3/NifH4	0.4720	0.1290	0.1916	13.3955
NifH3/NifH5	0.3376	0.1154	0.2124	8.5551
NifH3/NifH6	0.2856	0.1528	0.4451	3.4938
NifH4/NifH5	0.0584	0.1291	0.2847	0.2047
NifH4/NifH6	0.2352	0.1343	0.5115	3.0680
NifH5/NifH6	0.3048	0.1115	0.5097	7.4782

θ ML represent the maximum likelihood estimate for theta, coefficient of functional divergence; α ML represent the maximum likelihood estimate for alpha, the gamma shape parameter for rate variation among sites; SE θ represent standard error of the estimate theta and LRT θ represent 2 log-likelihood-ratio against the null hypothesis of θ=0.

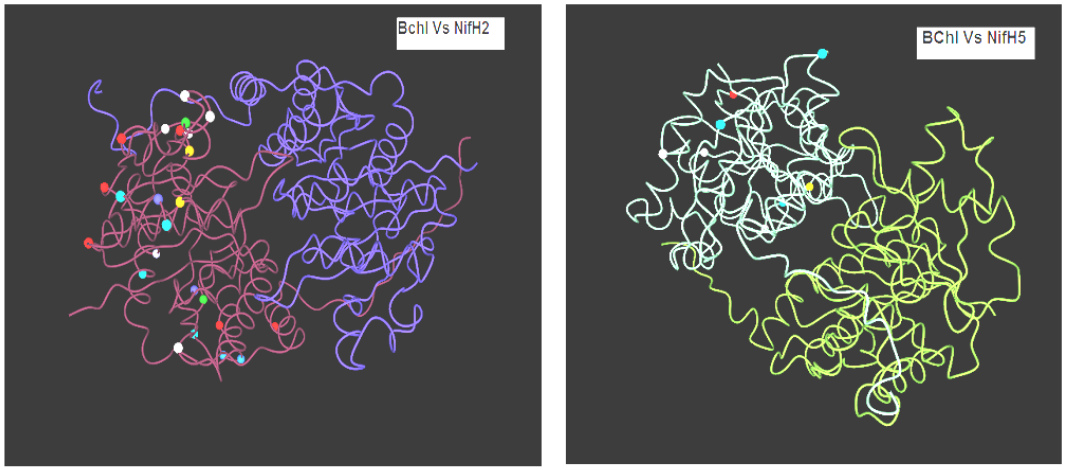
comprises of NifH sequences from Cyanobacteria except one from *Anabaena* whereas NifH2 subcluster has sequences from nitrogen-fixing actinobacteria like *Frankia*. NifH3&4 subcluster include legume associated symbiotic bacteria along with *Beijerinckia* and *Azospirillum* which which are crop rhizospheric bacteria. Sub cluster 5 & 6 consist free-living diazotrophs and anaerobic clostridia group. BchL subcluster include sequences of protochlorophyllide reductase from various anoxygenic phototrophic bacteria. This phylogeny was further used to detect site-specific altered functional constraints by comparing variability in evolutionary rates.

Table 4.9 shows the coefficient of functional divergence (θ) in various pair-wise comparisons. In all the cases, $\theta>0$ with $p<0.05$ indicating that likelihood ratio test implemented in detecting variability in evolutionary rates is found to be significant. It points towards rejection of null hypothesis i.e. $\theta=0$ signifying that evolutionary rate of the amino acid residues has significantly shifted between gene



A B

Figure 4.36 (A) and (B). Site specific profiles for evolutionary rate changes in the NifH/ Bchl protein family. The posterior probabilities of functional divergence for NifH2, NifH5 and Bchl were obtained with Diverge. Individual cut-off values for each comparison are marked with red horizontal lines.



A B

Figure 4.37 (A) and (B). Residues with predicted functional divergence between NifH/Bchl subclusters are mapped onto 3D structure. The highlighted residues represent the position of functionally divergent sites

clusters. Thus, Type I functional divergence analysis reveals *bchL*/*nif* genes are significantly functionally divergent from each other, owing to the evolutionary rate and/or property differences at some amino acid sites. Gene duplication could be the possible reason behind this site-specific rate shift in the protein family. There have been earlier reports of in-tandem gene duplication in *nif*D and *nif*K giving rise to the functional components of the enzyme (Fani *et al.*, 2000). Besides this, alternative supplemental forms of nitrogenase also been thought to have developed by gene duplications (Boyd *et al.*, 2011). Hence, it seems our observation regarding functional

divergence due to gene duplication in NifH/Bchl family is in line with that of earlier reports.

Further analysis was subsequently focused on pair wise comparison of the BchL/NifH2 and BchL/NifH5 sub-cluster. We had chosen these subclusters as they include members of an ancient symbiotic diazotroph and those in free-living condition. Amino acids residues responsible for functional divergence were identified using site-specific profiles (Figure 4.36 A&B) in combination with suitable cut -off-values derived from the posterior probability of each comparison. The cutoff value for the posterior probability was determined by consecutively eliminating the highest scoring residues until the co-efficient dropped to zero. Residues predicted to be functionally divergent were mapped

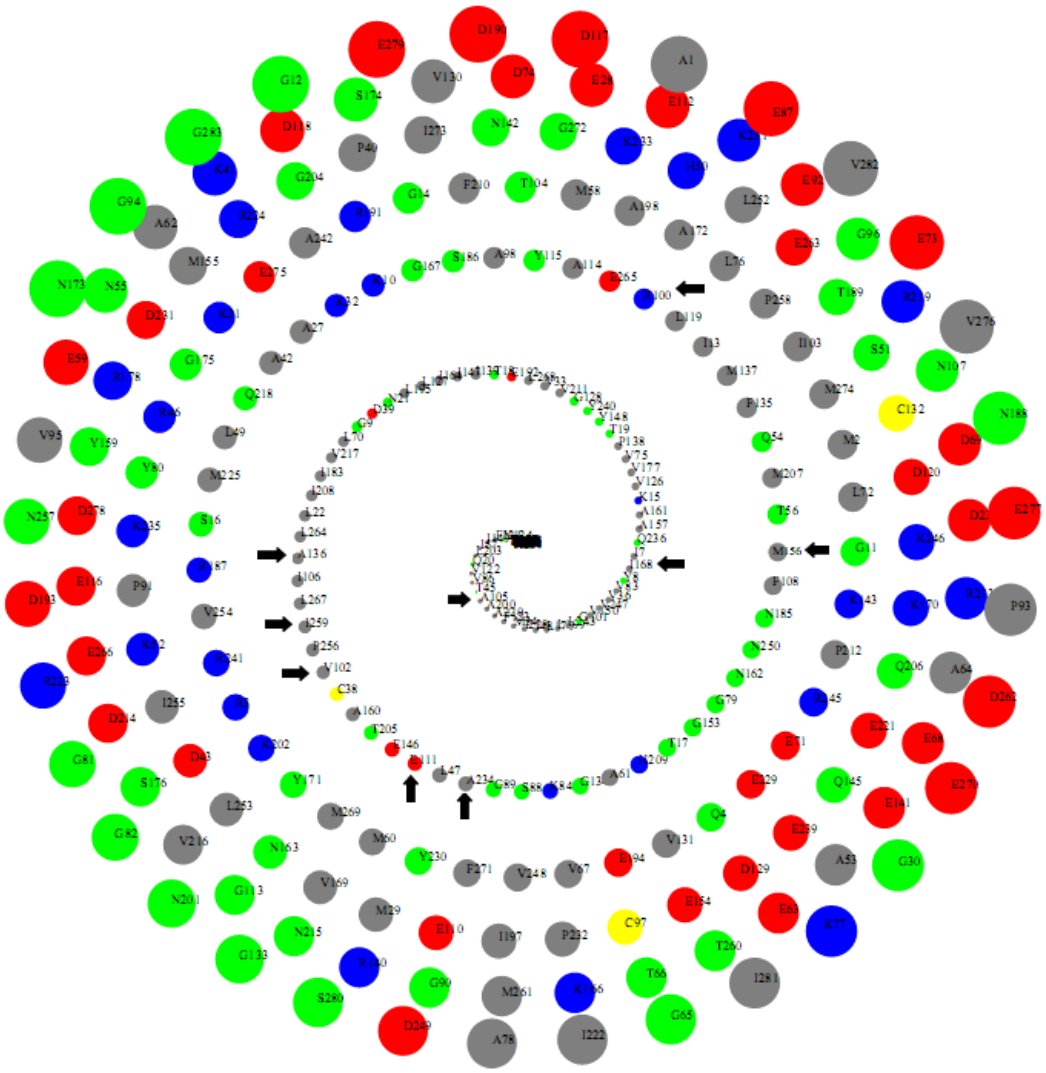


Figure 4.38 Solvent accessibility plot of the monomer of NifH protein. The black arrows represent the position of the functionally divergent residues with low solvent accessibility.

onto the three dimensional model of nitrogenase iron protein (Figure 4.37 A&B). These sites have varying evolutionary conservation among member genes

In the BchL/NifH2 comparison, positions of the functionally divergent residues are: 100, 102,104, 105, 109,111, 113, 122,136, 154, 156, 157, 161, 168, 199 , 227, 234, 248, 252, 253, 257, 259, 290, 293, 296, 325. Eight residues (having position 102, 104, 109, 136, 168, 224, 257, 296) were found to be functionally divergent while comparing between BchL and NifH5 subcluster. The arrangement of these residues in the 3D structure of NifH protein with relevance to their solvent accessibility was further investigated. The spiral plot (Figure 4.38) reveals the position of the residues according to the accessible surface area. From the plot, it can be seen that the functionally divergent residues particularly those near the position 100, 102, 105, 111, 136, 156, 168, 234, 259 are situated in the inner ring of the spiral. These residues are actually buried residues with relatively low solvent accessibility. Hence the residues are involved in maintaining the correct three-dimension structural framework of the protein.

4.8.2 Selective pressure among the amino-acid sites in the NifH family:

Functional divergence among proteins is often assumed to be strongly influenced by natural selection particularly positive selection. Presence of positive selection at amino acid residues was tested by implementing the site specific models in CODEML program. Likelihood rate test (LRT) performed between model M7 (beta) and M8 (beta and ω) divulges that M8 model is significantly favored. Positively selected sites were identified under M8 model using a Bayesian method. Eighteen amino acid sites inferred to under the influence of positive selection. Amongst them the sites with high posterior probabilities i.e. Bayes Empirical Bayes (BEB) analysis > 0.95 are Met-28, Glu-73, Gly-76, Arg-79, Tyr-180, Lys-188, Glu -189, Cys- 221, Glu- 222, Glu- 225, Glu-229 and Arg- 232. Some of these residues particularly those near the C-terminus have been known to mediate inter-subunit interactions. The residues in the c-terminus region wrap around the body of the opposing subunit and enhance the overall stabilization of the NifH dimer (Schlessman *et al.*, 1998). Therefore, Maximum likelihood (ML) tests implemented for investigating

selective pressure detected positive selection in amino acids involved in inter-subunit interactions in NifH/Bchl family. This indicates that positive selection has affected the evolution of these proteins

4.9 Statistical coupling analysis (SCA) of NifH/Bchl protein family:

NifH/BchL protein family includes members which are involved in ATP dependent transfer of electrons from a reductant, such as ferredoxin, via a metallic cluster. Although they share considerable structural similarities but they show broad range in terms of substrate specificities and environments within which they operate. The SCA calculation carried out by Matlab script for this family of protein illustrates an interesting case - sequences are distributed into distinct subfamilies, but this subfamily architecture is irrelevant to the pattern of positional correlations. Indeed, there is one apparent sector that seems to be a global property of members of this family regardless of the subfamily architecture.

First the sequence similarities analysis was carried in this family. The Figure 4.30 (A) shows a histogram of the similarities between pairs of sequences. The histogram depicts a multi-peaked distribution with groups of sequence showing mean pairwise identities between sequences ranging from 40% to greater than 60%. This pattern suggests the existence of multiple subfamilies. This assertion was further examined by direct visualization of the sequence similarity (S) matrix (Figure 4.39 B). The similarity matrix again indicates the presence of several subfamilies of sequences. The subfamily structure in this alignment indicates towards specific sectors related to various aspects of functional divergence between subfamilies.

The statistical coupling analysis (SCA) approach (Lockless and Ranganathan, 1999) was further utilized to examine the contribution of correlations to sequence conservation. A conservation weighted covariance matrix between all sequence positions in the NifH/Bchl protein family was computed with the help of Matlab script. This led to the identification of quasi-independent group of correlated amino acids in the member proteins. These structural elements are known as “sectors”. A single sector comprising of 72 amino acids have been identified in the NifH/ Bchl protein family. Figure 4.40 depicts the organization of the sector with respect to the three dimensional

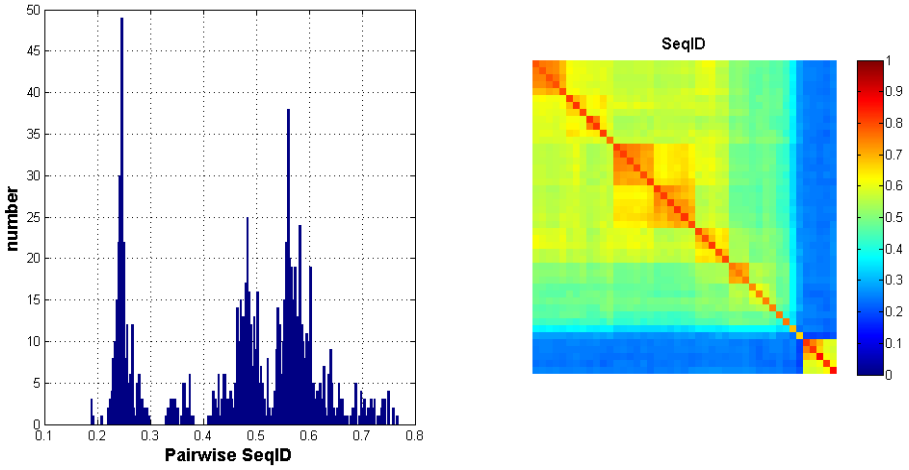


Figure 4.39 (A) Histogram of the similarities between pairs of sequences in the NifH/Bchl protein family ;(B) sequence similarity (S) matrix

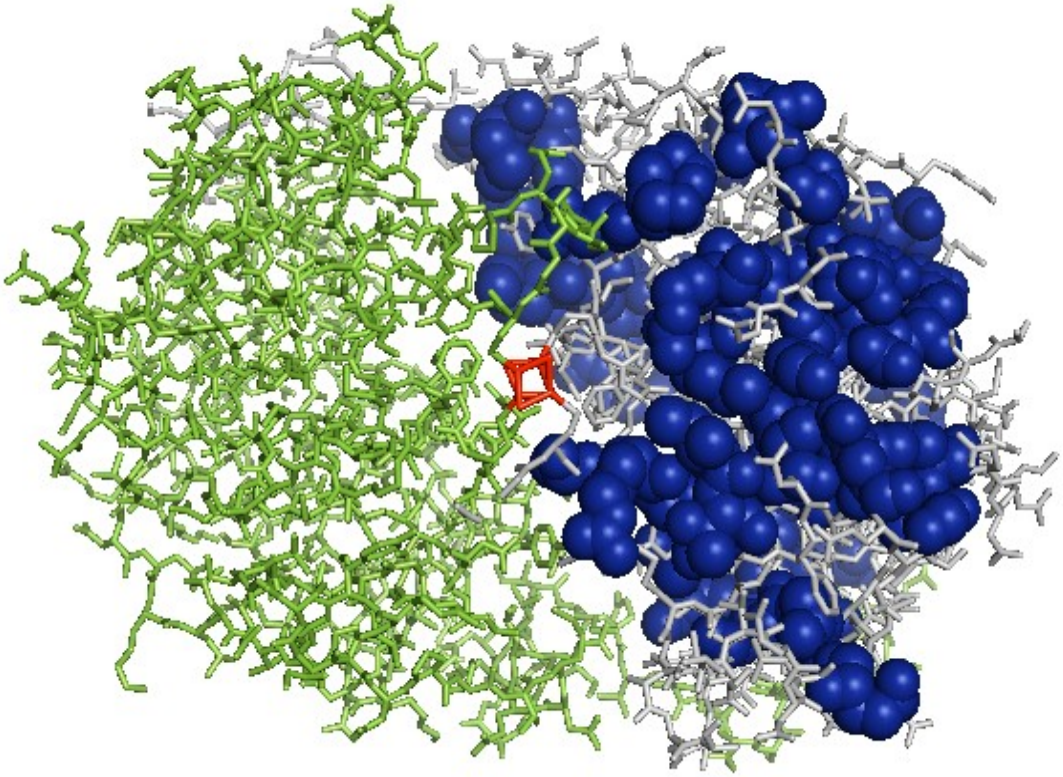


Figure 4.40 Residues comprising of the sector displayed in space filling representation on the tertiary structure of nitrogenase iron protein (pdb id 1G5P)

protein structure of NifH protein from *Azotobacter vindelandii* (pdb id 1G5P). The sector seems to be a network of interconnected residues. The position analysis of sector (blue) reveals that it forms a network of amino acids within the core domain that links the ligand-binding pocket to the allosteric surface sites. Thus, sectors represent the structural organization within proteins

reflecting, at least in part, the functional interactions between amino acid residues that underlies conserved biological properties. Sectors comprised of groups of coevolving amino acids that has a distinct functional role, and constitutes an independent mode of sequence divergence in the protein family. Thus, SCA yields one sector in the NifH/ BchL protein family that is a homogeneous property of sequences comprising the alignment. This is the case despite considerable heterogeneity of sequence divergence amongst members of the protein family

Conclusion

As we stride towards 21st Century, sustainable agriculture has become a necessity as well as reality. Biological nitrogen fixation forms an integral part of sustainable agriculture and therefore research in this sector has become the need of the moment. Research in BNF is not something totally out of box concept and has its beginning in early human civilization. But this work adds some new dimension to the on-going saga of biological nitrogen fixation. This is an attempt to investigate the nitrogen fixing machinery of symbiotic diazotroph and cyanobacteria in three levels i.e codon level, amino acid level and 3D protein structure level with the aid of various *in-silico* tools. Molecular modeling of protein and dynamic simulation provided a glimpse of the motion of the protein in *in-vivo* conditions –a feat which unfeasible to achieve in actual experimental conditions. Alternative phylogeny and detection of functional divergence of these proteins has been the other highlights of this work.

The major findings obtained from the study are summarized below:

- Synonymous codon usage analysis in the studied diazotrophs genomes revealed widespread variation at the interspecific and intraspecific levels. The results showed considerable degree of heterogeneity exists. Interplay of different factors like mutational pressure, translational efficiency, GC3 compositional bias etc. manipulated codon usage variation.
- It was found that AT-rich genomes show a bias at the third codon position with a strong preference towards bases Adenine and Thyamine. Reverse trend was found in the GC rich genomes, which frequently use codons that tend to end in either Cytosine or Guanine.
- Correspondence analysis based on amino acid composition of proteomes revealed striking

signatures, with sharp segregations within the diazotroph based on their lifestyle preferences.

- Analysis of rate of sequence evolution disclosed that most of the genes including the nitrogen fixation related genes (NFGs) in diazotroph genomes are under purifying selection. But the evolutionary rate of NFGs (in terms of Ka/Ks ratio) amongst *Frankia* stains is significantly different from those observed in other diazotrophs pointing towards relaxed purifying selection. Amongst the *nif*HDK genes, *nif*H was found to evolve slowly than others, indicating stronger stabilizing selection.
- The comparative genome analysis of fifteen Rhizobiales genomes with standardized gene findings helped in identification of a core and pangenome. The pangenome of rhizobiales turn out to be a "closed" one similar to those observed previously in many pathogens.
- Comparative genomics of *Frankia* strains showed that strains shared lower gene homology amongst themselves and gene clustering analysis further revealed that the strains are experiencing genome variation probably due to the adaptive response to their bio geographical positions.
- The 3D model of the NifH protein helped in identification of functionally important region of the protein. The molecular dynamics trajectory analysis involving the protein model revealed that functionally important motifs have stable secondary structure as well as tertiary arrangement during the simulation. The dimeric form of protein was found to quite stable in aqueous condition in comparison to the monomeric form possibly due to the wrapping of c-terminal ends.
- *In-silico* docking experiment involving the NifH protein highlighted the importance of Arginine 102 residue as covalent modification site in context of post-translational regulation of the nitrogenase enzyme.
- Nucleotide triplet phylogeny of *nif*HDK genes based on condensed matrix method advocated polyphyletic origin of these genes along with horizontal gene transfer and gene duplication as the major forces shaping the evolution.
- The structure based phylogeny disclosed that NifH protein and photosynthesis related light

independent protochlorophyllidae reducatse (BchL) shared a common ancestral protein. It suggested that NifH and ChlL could have evolved from a MinD like ancestor (1HYQ), a bacterial cell division regulator protein which itself evolved parallelly with Soj chromosome segregation protein (2BEK).

- Type1 functional divergence analysis lend support to the fact that functional evolution are marked by altered site-specific selective constrains possibly due to gene duplication. Critical amino acid residues relevant for distinct functional properties have also been identified. Positive selection pressure was detected in amino acid residues involved in subunit interactions.
- The SCA calculation carried out for the members of NifH/Bchl protein family illustrated that sequences are distributed into distinct subfamilies, but the subfamily architecture is irrelevant to the pattern of positional correlations. There is one apparent sector that seems to be a global property of members of this family regardless of the subfamily architecture.

Bibliography

Adams D (2000). Heterocyst formation in cyanobacteria. *Curr Opin Microbiol* **3**: 618 -624.

Agarwal G, Rajavel M, Gopal B and Srinivasan N (2009). Structure-based phylogeny as a diagnostic for functional characterization of proteins with a cupin fold. *PLoS One* **4**: e5736.

Alloisio N, Queiroux C, Fournier P, Pujic P, Normand P*, et al.* (2010). The Frankia alni symbiotic transcriptome. *Mol Plant Microbe Interact* **23**: 593-607.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403 -10.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z*, et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-402.

Amadou C, Pascal G, Mangenot S, Glew M, Bontemps C*, et al.* (2008). Genome sequence of the beta-rhizobium Cupriavidus taiwanensis and comparative genomics of rhizobia. *Genome Res* **18**: 1472-83.

Ampe F, Kiss E, Sabourdy F and Batut J (2003). Transcriptome analysis of Sinorhizobium meliloti during symbiosis. *Genome Biol* **4**: R15.

Anbar AD and Knoll AH (2002). Proterozoic ocean chemistry and evolution: a bioinorganic bridge? *Science* **297**: 1137-42.

Anthamatten D and Hennecke H (1991). The regulatory status of the fixL- and fixJ-like genes in Bradyrhizobium japonicum may be different from that in Rhizobium meliloti. *Mol Gen Genet* **225**: 38-48.

Appleby CA (1984). Leghemoglobin and Rhizobium respiration. *Annu Rev Plant Physiol* **33**: 443-478.

Ardourel M, Lortet G, Maillet F, Roche P, Truchet G*, et al.* (1995). In Rhizobium meliloti, the operon associated with the nod box n5 comprises nodL, noeA and noeB, three host-range genes specifically required for the nodulation of particular Medicago species. *Mol Microbiol* **17**: 687-99.

Arnold W, Rump A, Klipp W, Priefer UB and Puhler A (1988). Nucleotide sequence of a 24,206-base-pair DNA fragment carrying the entire nitrogen fixation gene cluster of Klebsiella pneumoniae. *J Mol Biol* **203**: 715-38.

Arp DJ (2000). The nitrogen cycle. In Prokaryotic Nitrogen Fixation. E. W. Triplett. Wymondham, Horizon Scientific Press**:** 1-14.

Austin S, Buck M, Cannon W, Eydmann T and Dixon R (1994). Purification and in vitro activities of the native nitrogen fixation control proteins NifA and NifL. *J Bacteriol* **176**: 3460-5.

Bachega JF, Navarro MV, Bleicher L, Bortoleto-Bugs RK, Dive D*, et al.* (2009). Systematic structural studies of iron superoxide dismutases from human parasites and a statistical coupling analysis of metal binding specificity. *Proteins* **77**: 26-37.

Baker W, van den Broek A, Camon E, Hingamp P, Sterk P*, et al.* (2000). The EMBL nucleotide sequence database. *Nucleic Acids Res* **28**: 19-23.

Balaji S and Srinivasan N (2001). Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. *Protein Eng* **14**: 219-26.

Balaji S and Srinivasan N (2007). Comparison of sequence-based and structure-based

phylogenetic trees of homologous proteins: Inferences on protein evolution. *J Biosci* **32**: 83-96.

Baldani JI and Baldani VL (2005). History on the biological nitrogen fixation research in graminaceous plants: special emphasis on the Brazilian experience. *An Acad Bras Cienc* **77**: 549-79.

Banci L and Comba P (1997). Molecular modeling and dynamics of bioinorganic systems, Kluwer Academic Pub.

Banci L (2003). Molecular dynamics simulations of metalloproteins. *Curr Opin Chem Biol* **7**: 143-9.

Batut J, Santero E and Kustu S (1991). In vitro activity of the nitrogen fixation regulatory protein FIXJ from Rhizobium meliloti. *J Bacteriol* **173**: 5914-7.

Batut J and Boistard P (1994). Oxygen control in Rhizobium. *Anton Van Leeuwen* **66**: 129 -50.

Bayly CI, Cieplak P, Cornell W and Kollman PA (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem* **97**: 10269-10280.

Becker A, Barnett MJ, Capela D, Dondrup M, Kamp PB*, et al.* (2009). A portal for rhizobial genomes: RhizoGATE integrates a Sinorhizobium meliloti genome annotation update with postgenome data. *J Biotechnol* **140**: 45-50.

Bell CJ, Dixon RA, Farmer AD, Flores R, Inman J*, et al.* (2001). The Medicago Genome Initiative: a model legume database. *Nucleic Acids Res* **29**: 114-7.

Bennetzen JL and Hall BD (1982). Codon selection in yeast. *J Biol Chem* **257**: 3026-31.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA*, et al.* (2000). GenBank. *Nucleic Acids Res* **28**: 15-8.

Benson DR and Silvester WB (1993). Biology of Frankia strains, actinomycete symbionts of actinorhizal plants. *Microbiol Rev* **57**: 293-319.

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

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

Berges H, Lauber E, Liebe C, Batut J, Kahn D*, et al.* (2003). Development of Sinorhizobium meliloti pilot macroarrays for transcriptome analysis. *Appl Environ Microbiol* **69**: 1214-9.

Berman-Frank I, Lundgren P and Falkowski P (2003). Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria. *Res Microbiol* **154**: 157-64.

Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD*, et al.* (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem* **80**: 319-24.

Berry AM, Harriott OT, Moreau RA, Osman SF, Benson DR*, et al.* (1993). Hopanoid lipids compose the Frankia vesicle envelope, presumptive barrier of oxygen diffusion to nitrogenase. *Proc Natl Acad Sci U S A* **90**: 6091-4.

Binnewies TT, Hallin PF, Staerfeldt HH and Ussery DW (2005). Genome Update: proteome comparisons. *Microbiology* **151**: 1-4.

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

Bishop PE and Premakumar R (1992). Alternative nitrogen fixation systems. In G. Stacey, R. H. Burris, H. J. Evans (Eds.), Biological nitrogen fixation, New York: Chapman and Hall.

Black M, Moolhuijzen P, Chapman B, Barrero R, Howieson J*, et al.* (2012). The Genetics of Symbiotic Nitrogen Fixation: Comparative Genomics of 14 Rhizobia Strains by Resolution of Protein Clusters. *Genes* **3**: 138-166.

Bohlool BB, Ladha JK, Garrity DP and George T (1992). Biological nitrogen fixation for sustainable agriculture: A perspective. *Plant and Soil* **141**: 1-11.

Bohme H and Haselkorn R (1988). Molecular cloning and nucleotide sequence analysis of the gene coding for heterocyst ferredoxin from the cyanobacterium Anabaena sp. strain PCC 7120. *Mol Gen Genet* **214**: 278-85.

Bolin JT, Ronco AE, Morgan TV, Mortenson LE and Xuong NH (1993a). The unusual metal clusters of nitrogenase: structural features revealed by x-ray anomalous diffraction studies of the MoFe protein from Clostridium pasteurianum. *Proc Natl Acad Sci U S A* **90**: 1078-82.

Bolin TJ, Campobasso N, Muchmore SW, Morgan TV and Mortenson LE (1993b). The structure and environment of the metal clusters in the MoFe protein from Clostridium pasteurianum. Molybdenum Enzymes,Cofactors, and Model Systems. E. I. Stiefel, D. Coucouvanis and W. E. Newton. Washington DC, American Chemical Society**:** 186–195.

Bonaldi K, Gourion B, Fardoux J, Hannibal L, Cartieaux F*, et al.* (2010). Large-scale transposon mutagenesis of photosynthetic Bradyrhizobium sp. strain ORS278 reveals new genetic loci putatively important for nod-independent symbiosis with Aeschynomene indica. *Mol Plant Microbe Interact* **23**: 760-70.

Bormann BT, Bormann FH, Bowden WB, Piece RS, Hamburg SP*, et al.* (1993). Rapid N 2 Fixation in Pines, Alder, and Locust: Evidence From the Sandbox Ecosystems Study. *Ecology*583-598.

Borthakur D, Basche M, Buikema WJ, Borthakur PB and Haselkorn R (1990). Expression, nucleotide sequence and mutational analysis of two open reading frames in the nif gene region of Anabaena sp. strain PCC7120. *Mol Gen Genet* **221**: 227-34.

Boyd ES, Hamilton TL and Peters JW (2011). An alternative path for the evolution of biological nitrogen fixation. *Front Microbiol* **2**: 205.

Breitling R, Laubner D and Adamski J (2001). Structure-based phylogenetic analysis of short-chain alcohol dehydrogenases and reclassification of the 17beta-hydroxysteroid dehydrogenase family. *Mol Biol Evol* **18**: 2154-61.

Brewin NJ (1993). The Rhizobium-legume symbiosis: plant morphogenesis in a nodule. *Semin Cell Biol* **4**: 149-56.

Brigle KE, Weiss MC, Newton WE and Dean DR (1987). Products of the iron-molybdenum cofactor-specific biosynthetic genes, nifE and nifN, are structurally homologous to the products of the nitrogenase molybdenum-iron protein genes, nifD and nifK. *J Bacteriol* **169**: 1547-53.

Brito B, Martinez M, Fernandez D, Rey L, Cabrera E*, et al.* (1997). Hydrogenase genes from Rhizobium leguminosarum bv. viciae are controlled by the nitrogen fixation regulatory protein nifA. *Proc Natl Acad Sci U S A* **94**: 6019-24.

Brooks B and Karplus M (1983). Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci U S A* **80**: 6571-5.

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

Burgess BK and Lowe DJ (1996). Mechanism of Molybdenum Nitrogenase. *Chem Rev* **96**: 2983-3012.

Burke DH, Hearst JE and Sidow A (1993). Early evolution of photosynthesis: clues from nitrogenase and chlorophyll iron proteins. *Proc Natl Acad Sci U S A* **90**: 7134-8.

Burns RC and Hardy RW (1975). Nitrogen fixation in bacteria and higher plants. *Mol Biol Biochem Biophys*1-189.

Bursey EH and Burgess BK (1998). Characterization of a variant iron protein of nitrogenase that is impaired in its ability to adopt the MgATP-induced conformational change. *J Biol Chem* **273**: 16927-34.

Cantera JJ, Kawasaki H and Seki T (2004). The nitrogen-fixing gene (nifH) of Rhodopseudomonas palustris: a case of lateral gene transfer? *Microbiology* **150**: 2237-46.

Capone DG (2001). Marine nitrogen fixation: what's the fuss? *Curr Opin Microbiol* **4**: 341-8.

Capriotti E, Fariselli P and Casadio R (2005). I -Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res* **33**: W306-W310.

Carvalho FM, Souza RC, Barcellos FG, Hungria M and Vasconcelos AT (2010). Genomic and evolutionary comparisons of diazotrophic and pathogenic bacteria of the order Rhizobiales. *BMC Microbiol* **10**: 37.

Case DA and Karplus M (1979). Dynamics of ligand binding to heme proteins. *J Mol Biol* **132**: 343-68.

Centeno NB, Planas-Iglesias J and Oliva B (2005). Comparative modelling of protein structure and its impact on microbial cell factories. *Microb Cell Fact* **4**: 20.

Ceremonie H, Cournoyer B, Maillet F, Normand P and Fernandez MP (1998).

Genetic complementation of rhizobial nod mutants with Frankia DNA: artifact or reality? *Mol Gen Genet* **260**: 115-119.

Chan MK, Kim J and Rees DC (1993). The nitrogenase FeMo-cofactor and P-cluster pair: 2.2 A resolution structures. *Science* **260**: 792-4.

Chang G, Roth CB, Reyes CL, Pornillos O, Chen YJ*, et al.* (2006). Retraction. *Science* **314**: 1875.

Chen H, Higgins J, Oresnik IJ, Hynes MF, Natera S*, et al.* (2000). Proteome analysis demonstrates complex replicon and luteolin interactions in pSyma-cured derivatives of Sinorhizobium meliloti strain 2011. *Electrophoresis* **21**: 3833-42.

Chen WX, Yan GH and Li JL (1988). Numerical taxonomic study of fast-growing soybean rhizobia and a proposal that Rhizobium fredii be assigned to Sinorhizobium gen. nov. *Int J Syst Bacteriol.* **38**: 392-397.

Chisnell JR, Premakumar R and Bishop PE (1988). Purification of a second alternative nitrogenase from a nifHDK deletion strain of Azotobacter vinelandii. *J Bacteriol* **170**: 27-33.

Chiu H, Peters JW, Lanzilotta WN, Ryle MJ, Seefeldt LC*, et al.* (2001). MgATP-Bound and nucleotide-free structures of a nitrogenase protein complex between the Leu 127 Delta-Fe-protein and the MoFe-protein. *Biochemistry* **40**: 641-50.

Choi JH, Jung HY, Kim HS and Cho HG (2000). PhyloDraw: a phylogenetic tree drawing system. *Bioinformatics* **16**: 1056-8.

Chothia C, Lesk AM, Levitt M, Amit AG, Mariuzza RA*, et al.* (1986). The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure. *Science* **233**: 755-8.

Cleveland CC, Townsend AR, Schimel DS, Fisher H, Howarth RW*, et al.* (1999). Global patterns of terrestrial biological nitrogen (N2) fixation in natural ecosystems. *Global Biogeochem. Cycles* **13**: 623-645.

Colebatch G, Kloska S, Trevaskis B, Freund S, Altmann T*, et al.* (2002). Novel aspects of symbiotic nitrogen fixation uncovered by transcript profiling with cDNA arrays. *Mol Plant Microbe Interact* **15**: 411-20.

Colovos C and Yeates TO (1993). Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci* **2**: 1511-9.

Comba P and Remenyi R (2002). A new molecular mechanics force field for the oxidized form of blue copper proteins. *J Comput Chem* **23**: 697-705.

Covitz PA, Smith LS and Long SR (1998). Expressed sequence tags from a root-hair-enriched medicago truncatula cDNA library. *Plant Physiol* **117**: 1325-32.

Crossman LC, Castillo-Ramirez S, McAnnula C, Lozano L, Vernikos GS*, et al.* (2008). A common genomic framework for a diverse assembly of plasmids in the symbiotic nitrogen fixing bacteria. *PLoS One* **3**: e2567.

D'Auria G, Jimenez-Hernandez N, Peris-Bondia F, Moya A and Latorre A Legionella pneumophila pangenome reveals strain-specific virulence factors. *BMC Genomics* **11**: 181.

Dawson JO (1983). Dinitrogen fixation in forest ecosystems. *Can J. Microbiol* **29**:979-992

Dean DR and Jacobson MR (1992). Biochemical genetics of nitrogenase. In G. Stacey, R. H. Burris, H. J. Evans (Eds.), Biological nitrogen fixation, New York: Chapman and Hall.

Dedysh SN, Ricke P and Liesack W (2004). NifH and NifD phylogenies: an evolutionary basis for understanding nitrogen fixation capabilities of methanotrophic bacteria. *Microbiology* **150**: 1301-13.

Delgado MJ, Bedmar EJ and Downie JA (1998). Genes involved in the formation and assembly of rhizobial cytochromes and their role in symbiotic nitrogen fixation. *Adv Microb Physiol* **40**: 191-231.

Denarie J, Debelle F and Prome JC (1996). Rhizobium lipo-chitooligosaccharide nodulation factors: signaling molecules mediating recognition and morphogenesis. *Annu Rev Biochem* **65**: 503-35.

Deng X, Phillippy AM, Li Z, Salzberg SL and Zhang W (2010). Probing the pan-genome of Listeria monocytogenes: new insights into intraspecific niche expansion and genomic diversification. *BMC Genomics* **11**: 500.

Dermitzakis ET and Clark AG (2001). Differential selection after duplication in mammalian developmental genes. *Mol Biol Evol* **18**: 557-62.

Dixon R and Kahn D (2004). Genetic regulation of biological nitrogen fixation. *Nat Rev Microbiol* **2**: 621-31.

Dixon R, Eady RR, Espin G, Hill S, Iaccarino M*, et al.* (1980). Analysis of regulation of Klebsiella pneumoniae nitrogen fixation (nif) gene cluster with gene fusions. *Nature* **286**: 128-32.

Dobereiner J (1976). Plant genotype effects on nitrogen fixation in grasses. *Basic Life Sci* **8**: 325-34.

Dodson GG, Lane DP and Verma CS (2008). Molecular simulations of protein dynamics: new windows on mechanisms in biology. *EMBO Rep* **9**: 144-50.

Dommergues YR (1997). Contribution of actinorhizal plants to tropical soil productivity and rehabilitation. *Soil Biol Biochem* **29**: 931-941.

dos Reis M, Wernisch L and 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**: 6976-85.

dos Santos ES, Gritta DS, Taft CA, Almeida PF and Ramos-de-Souza E (2009). Molecular dynamics simulation of the adenylylsulphate reductase from hyperthermophilic Archaeoglobus fulgidus. *Mol Simulat* **36**: 199-203.

Dos Santos PC, Fang Z, Mason SW, Setubal JC and Dixon R (2012). Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics* **13**: 162.

Downie JA (2011). The roles of extracellular proteins, polysaccharides and signals in the interactions of rhizobia with legume roots. *FEMS Microbiol Rev* **34**: 150-70.

Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y*, et al.* (2006). CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res* **34**: W116-8.

Eady RR (1996). Structure-Function Relationships of Alternative Nitrogenases. *Chem Rev* **96**: 3013-3030.

Eady RR and Postgate JR (1974). Nitrogenase. *Nature* **249**: 805-10.

Einsle O, Tezcan FA, Andrade SL, Schmid B, Yoshida M*, et al.* (2002). Nitrogenase MoFe-protein at 1.16 A resolution: a central ligand in the FeMo-cofactor. *Science* **297**: 1696-700.

Eisenberg D, Luthy R and Bowie JU (1997). VERIFY3D: Assessment of protein models with three-dimensional profiles. *Method enzymol* **277**: 396-404.

Elmerich C (1984). Molecular biology and ecology of diazotrophs associated with non -leguminous plants. *Nat Biotechnol* **2**: 967-978.

Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA*, et al.* (2003). Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res* **31**: 3375-80.

Falkowski PG (1997). Evolution of the nitrogen cycle and its influence on the biological sequestration of CO2 in the ocean. *Nature* **387**: 272-275.

Fani R, Gallo R and Lio P (2000). Molecular evolution of nitrogen fixation: the evolutionary history of the nifD, nifK, nifE, and nifN genes. *J Mol Evol* **51**: 1-11.

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

Fenstermacher D (2005). Introduction to bioinformatics. *J Am Soc Inf* **56**: 440-446.

Fischer HM (1994). Genetic regulation of nitrogen fixation in rhizobia. *Microbiol Rev* **58**: 352-86.

Fitch WM and Margoliash E (1967). Construction of phylogenetic trees. *Science* **155**: 279-84.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF*, et al.* (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**: 496-512.

Fogg GE (1978). Nitrogen Fixation in the Oceans. *Ecological Bulletins*11-19.

Fox T and Kollman PA (1998). Application of the RESP methodology in the parametrization of organic solvents. *J Phys Chem B* **102**: 8070-8079.

Franche C, Lindstrom K and Elmerich C (2009). Nitrogen-fixing bacteria associated with leguminous and non-leguminous plants. *Plant and soil* **321**: 35-59.

Freiberg C, Perret X, Broughton WJ and Rosenthal A (1996). Sequencing the 500-kb GC-rich symbiotic replicon of Rhizobium sp. NGR234 using dye terminators and a thermostable "sequenase": a beginning. *Genome Res* **6**:

590-600.

Friis C, Wassenaar TM, Javed MA, Snipen L, Lagesen K, *et al.* (2010). Genomic characterization of Campylobacter jejuni strain M1. *PLoS One* **5**: e12253.

Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, *et al.* (2003). Gaussian 03, Revision C.02.

Fujita Y and Bauer CE (2000). Reconstitution of light-independent protochlorophyllide reductase from purified bchl and BchN-BchB subunits.. *J Biol Chem* **275**: 23583-8.

Fujita Y, Takahashi Y, Kohchi T, Ozeki H, Ohyama K, *et al.* (1989). Identification of a novel nifH-like (frxC) protein in chloroplasts of the liverwort Marchantia polymorpha. *Plant Mol Biol* **13**: 551-61.

Galardini M, Mengoni A, Brilli M, Pini F, Fioravanti A, *et al.* (2011). Exploring the symbiotic pangenome of the nitrogen-fixing bacterium Sinorhizobium meliloti. *BMC Genomics* **12**: 235.

Galibert F, Finan TM, Long SR, Puhler A, Abola P, *et al.* (2001). The composite genome of the legume symbiont Sinorhizobium meliloti. *Science* **293**: 668-72.

Garau G, Di Guilmi AM and Hall BG (2005). Structure-based phylogeny of the metallo-beta-lactamases. *Antimicrob Agents Chemother* **49**: 2778-84.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, *et al.* (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **31**: 3784-8.

Gatti D, Mitra B and Rosen BP (2000). Escherichia coli soft metal ion-translocating ATPases. *J Biol Chem* **275**: 34009-12.

Georgiadis MM, Komiya H, Chakrabarti P, Woo D, Kornuc JJ, *et al.* (1992). Crystallographic structure of the nitrogenase iron protein from Azotobacter vinelandii. *Science* **257**: 1653-9.

Geurts R and Bisseling T (2002). Rhizobium Nod factor perception and signalling. *Plant Cell* **14**: S239-S249.

Giller KE and Day JM (1985). Nitrogen fixation in the rhizosphere: significance in natural and agricultural systems. Ecological Interactions in Soil. Plants Microbes and Animals. A. H. Fitter. Oxford, Blackwell Scientific Publications.

Giraud E, Moulin L, Vallenet D, Barbe V, Cytryn E, *et al.* (2007). Legumes symbioses: absence of Nod genes in photosynthetic bradyrhizobia. *Science* **316**: 1307-12.

Glazer AN and Kechris KJ (2009). Conserved amino acid sequence features in the alpha subunits of MoFe, VFe, and FeFe nitrogenases. *PLoS One* **4**: e6136.

Golden JW, Robinson SJ and Haselkorn R (1985). Rearrangement of nitrogen fixation genes during heterocyst differentiation in the cyanobacterium Anabaena. *Nature* **314**: 419-23.

Gonzalez V, Bustos P, Ramirez-Romero MA, Medrano-Soto A, Salgado H, *et al.* (2003). The mosaic structure of the symbiotic plasmid of Rhizobium etli CFN42 and its relation to other symbiotic genome compartments. *Genome Biol* **4**: R36.

Gonzalez V, Santamaria RI, Bustos P, Hernandez-Gonzalez I, Medrano-Soto A, *et al.* (2006). The partitioned Rhizobium etli genome: genetic and metabolic redundancy in seven interacting replicons. *Proc Natl Acad Sci U S A* **103**: 3834-9.

Gribaldo S, Casane D, Lopez P and Philippe H (2003). Functional divergence prediction from evolutionary analysis: a case study of vertebrate hemoglobin. *Mol Biol Evol* **20**: 1754-9.

Grottesi A, Ceruso MA, Colosimo A and Di Nola A (2002). Molecular dynamics study of a hyperthermophilic and a mesophilic rubredoxin. *Proteins* **46**: 287-94.

Gtari M, Ghodhbane-Gtari F, Nouioui I, Beauchemin N and Tisa LS (2012). Phylogenetic perspectives of nitrogen-fixing actinobacteria. *Arch Microbiol* **194**: 3-11.

Gu X (1999). Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol* **16**: 1664-74.

Gu X (2001). Maximum-likelihood approach for gene family evolution under functional divergence. *Mol Biol Evol* **18**: 453-64.

Gubler M and Hennecke H (1986). fixA, B and C genes are essential for symbiotic and free-living, microaerobic nitrogen fixation. *FEBS letters* **200**: 186-192.

Guermeur Y (1997). Combinaison de classifieurs statistiques, Application à la prédiction de la structure secondaire des

protéines, Université Paris 6.

Guerreiro N, Djordjevic MA and Rolfe BG (1999). Proteome analysis of the model microsymbiont Sinorhizobium meliloti: isolation and characterisation of novel proteins. *Electrophoresis* **20**: 818-25.

Guindon S and Gascuel O (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696-704.

Guldner E, Desmarais E, Galtier N and Godelle B (2004). Molecular evolution of plant haemoglobin: two haemoglobin genes in nymphaeaceae Euryale ferox. *J Evol Biol* **17**: 48-54.

Gyorgyey J, Vaubert D, Jimenez-Zurdo JI, Charon C, Troussard L*, et al.* (2000). Analysis of Medicago truncatula nodule expressed sequence tags. *Mol Plant Microbe Interact* **13**: 62-71.

Hageman RV and Burris RH (1978). Nitrogenase and nitrogenase reductase associate and dissociate with each catalytic cycle. *Proc Natl Acad Sci U S A* **75**: 2699-702.

Halabi N, Rivoire O, Leibler S and Ranganathan R (2009). Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**: 774-86.

Hales BJ, Case EE, Morningstar JE, Dzeda MF and Mauterer LA (1986). Isolation of a new vanadium-containing nitrogenase from Azotobacter vinelandii. *Biochemistry* **25**: 7251-5.

Hartmann LS and Barnum SR (2010). Inferring the evolutionary history of Mo-dependent nitrogen fixation from phylogenetic studies of nifK and nifDK. *J Mol Evol* **71**: 70-85.

Haukka K, Lindstrom K and Young JP (1998). Three phylogenetic groups of nodA and nifH genes in Sinorhizobium and Mesorhizobium isolates from leguminous trees growing in Africa and Latin America. *Appl Environ Microbiol* **64**: 419-26.

Hennecke H, Kaluza K, Thöny B, Fuhrmann M, Ludwig W*, et al.* (1985). Concurrent evolution of nitrogenase genes and 16S rRNA in Rhizobium species and other nitrogen fixing bacteria. *Arch Microbiol* **142**: 342-348.

Henson BJ, Hesselbrock SM, Watson LE and Barnum SR (2004a). Molecular phylogeny of the heterocystous cyanobacteria (subsections IV and V) based on nifD. *Int J Syst Evol Microbiol* **54**: 493-7.

Henson BJ, Watson LE and Barnum SR (2004b). The evolutionary history of nitrogen fixation, as assessed by NifD. *J Mol Evol* **58**: 390-9.

Hesper B and Hogeweg P (1970). Bioinformatica: een werkconcept. *Kameleon* **1**: 28-29.

Hess B, Bekker H, Berendsen HJC and Fraaije JGEM (1997). LINCS: a linear constraint solver for molecular simulations. *J Comput Chem* **18**: 1463-1472.

Hiller NL, Janto B, Hogg JS, Boissy R, Yu S*, et al.* (2007). Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. *J Bacteriol* **189**: 8186-95.

Hirsch A (2009). Brief History of the Discovery of Nitrogen Fixing Organisms. *Available at Web site http://www. mcdb. ucla. edu/Research/Hirsch/imagesb/ HistoryDiscoveryN2fixing Organisms.*

Hirsch AM, McKhann HI, Reddy A, Liao J, Fang Y*, et al.* (1995). Assessing horizontal transfer of nifHDK genes in eubacteria: nucleotide sequence of nifK from Frankia strain HFPCcI3. *Mol Biol Evol* **12**: 16-27.

Hoffman BM, Dean DR and Seefeldt LC (2009). Climbing nitrogenase: toward a mechanism of enzymatic nitrogen fixation. *Acc Chem Res* **42**: 609-19.

Hoffman K and Kunze R (1971). Linear Algebra., Prentice-Hall, Englewood Cliffs, NJ.

Hogeweg P (2011). The roots of bioinformatics in theoretical biology. *PLoS Comput Biol* **7**: e1002021.

Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J*, et al.* (2007). Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol* **8**: R103.

Hollup SM, Salensminde G and Reuter N (2005). WEBnm@: a web application for normal mode analyses of proteins. *BMC Bioinformatics* **6**: 52.

Holm L and Sander C (1995). Dali: a network tool for protein structure comparison. *Trends Biochem Sci* **20**: 478-80.

Howard JB and Rees DC (1996). Structural Basis of Biological Nitrogen Fixation.

Chem Rev **96**: 2965-2982.

Hu Z, Saez C and Lutkenhaus J (2003). Recruitment of MinC, an inhibitor of Z-ring formation, to the membrane in Escherichia coli: role of MinD and MinE. *J Bacteriol* **185**: 196-203.

Hubbard TJ and Blundell TL (1987). Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng* **1**: 159-71.

Hurek T, Egener T and Reinhold-Hurek B (1997). Divergence in nitrogenases of Azoarcus spp., Proteobacteria of the beta subclass. *J Bacteriol* **179**: 4172-8.

Hurst LD (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet* **18**: 486.

Hurst LD and Pal C (2001). Evidence for purifying selection acting on silent sites in BRCA1. *Trends Genet* **17**: 62-5.

Huss-Danell K (1997). Actinorhizal symbioses and their N2 fixation. *New Phytol* **136**: 375 –405.

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

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

Imperial J, Ugalde RA, Shah VK and Brill WJ (1984). Role of the nifQ gene product in the incorporation of molybdenum into nitrogenase in Klebsiella pneumoniae. *J Bacteriol* **158**: 187-94.

Irisarri P, Gonnet S and Monza J (2001). Cyanobacteria in Uruguayan rice fields: diversity, nitrogen fixing ability and tolerance to herbicides and combined nitrogen. *J Biotechnol* **91**: 95-103.

Jacobsen A, Hendriksen RS, Aaresturp FM, Ussery DW and Friis C (2011). The Salmonella enterica pan-genome. *Microb Ecol* **62**: 487-504.

Jang SB, Seefeldt LC and Peters JW (2000). Insights into nucleotide signal transduction in nitrogenase: structure of an iron protein with MgADP bound. *Biochemistry* **39**: 14745-52.

Jarvis BD, Downer HL and Young JP (1992). Phylogeny of fast-growing soybean-nodulating rhizobia support synonymy of Sinorhizobium and Rhizobium and assignment to Rhizobium fredii. *Int J Syst Bacteriol* **42**: 93-6.

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

Jones DT (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**: 195-202.

Jurgensen MF, Graham RT, Larsen MJ and Harvey AE (1992). Clear-cutting, woody residue removal, and nonsymbiotic nitrogen fixation in forest soils of the Inland Pacific Northwest. *Can J Forest Res* **22**: 1172-1178.

Kabsch W and Sander C (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**: 2577-637.

Kaneko T, Nakamura Y, Sato S, Asamizu E, Kato T, *et al.* (2000). Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti (supplement). *DNA Res* **7**: 381-406.

Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, *et al.* (2002). Complete genomic sequence of nitrogen-fixing symbiotic bacterium Bradyrhizobium japonicum USDA110 (supplement). *DNA Res* **9**: 225-56.

Kaplan W and Littlejohn TG (2001). Swiss-PDB viewer (deep view). *Brief Bioinform* **2**: 195-197.

Kechris KJ, Lin JC, Bickel PJ and Glazer AN (2006). Quantitative exploration of the occurrence of lateral gene transfer by using nitrogen fixation genes as a case study. *Proc Natl Acad Sci U S A* **103**: 9584-9.

Kennedy IR and Islam N (2001). The current and potential contribution of asymbiotic nitrogen fixation to nitrogen requirements on farms: a review. *Aust J Exp Agr* **41**: 447 -457.

Kent HM, Ioannidis I, Gormal C, Smith BE and Buck M (1989). Site-directed mutagenesis of the Klebsiella pneumoniae nitrogenase. Effects of modifying conserved cysteine residues in the alpha- and beta-subunits. *Biochem J* **264**: 257-64.

Kessler PS, McLarnan J and Leigh JA (1997). Nitrogenase phylogeny and the molybdenum dependence of nitrogen fixation in Methanococcus maripaludis. *J*

bacterial. **179**: 541-543.

Kim CH, Newton WE and Dean DR (1995). Role of the MoFe protein alpha-subunit histidine-195 residue in FeMo-cofactor binding and nitrogenase catalysis. *Biochemistry* **34**: 2798-808.

Kim J and Rees DC (1992). Structural models for the metal centers in the nitrogenase molybdenum-iron protein. *Science* **257**: 1677-82.

Kim J, Woo D and Rees DC (1993). X-ray crystal structure of the nitrogenase molybdenum-iron protein from Clostridium pasteurianum at 3.0-A resolution. *Biochemistry* **32**: 7104-15.

Klipp W (2004). Genetics and regulation of nitrogen fixation in free-living bacteria. Dordrecht ; Boston, Kluwer Academic Publishers.

Knight RD, Freeland SJ and 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**: RESEARCH0010.

Kucho K, Hay AE and Normand P (2010). The determinants of the actinorhizal symbiosis. *Microbes Environ* **25**: 241-52.

Kuhn B, Jacobsen W, Christians U, Benet LZ and Kollman PA (2001). Metabolism of sirolimus and its derivative everolimus by cytochrome P450 3A4: insights from docking, molecular dynamics, and quantum chemical calculations. *J Med Chem* **44**: 2027-34.

Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T*, et al.* (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**: 3100-8.

Lahiri S, Pulakat L and Gavini N (2008). NifH: Structural and Mechanistic Similarities with Proteins Involved in Diverse Biological Processes. *Am J Biochem Biotechnol* **4**: 304-316.

Landgraf R, Fischer D and Eisenberg D (1999). Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng* **12**: 943-51.

Laskowski RA, Watson JD and Thornton JM (2005). ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* **33**: W89-93.

Latysheva N, Junker VL, Palmer WJ, Codd GA and Barker D (2012). The evolution of nitrogen fixation in cyanobacteria. *Bioinformatics* **28**: 603-606

Lee KB, De Backer P, Aono T, Liu CT, Suzuki S*, et al.* (2008). The genome of the versatile nitrogen fixer Azorhizobium caulinodans ORS571. *BMC Genomics* **9**: 271.

Lei S, Pulakat L and Gavini N (1999). Genetic analysis of nif regulatory genes by utilizing the yeast two-hybrid system detected formation of a NifL-NifA complex that is implicated in regulated expression of nif genes. *J bacteriol* **181**: 6535-6539.

Leigh GJ (2002). Nitrogen fixation at the millennium. Amsterdam ; London, Elsevier.

Lery LM, Bitar M, Costa MG, Rossle SC and Bisch PM (2010). Unraveling the molecular mechanisms of nitrogenase conformational protection against oxygen in diazotrophic bacteria. *BMC Genomics* **11 Suppl 5**: S7.

Levitt M and Sharon R (1988). Accurate simulation of protein dynamics in solution. *Proc Natl Acad Sci U S A* **85**: 7557-61.

Li W, Liu B, Yu L, Feng D, Wang H*, et al.* (2009a). Phylogenetic analysis, structural evolution and functional divergence of the 12-oxo-phytodienoate acid reductase gene family in plants. *BMC Evol Biol* **9**: 90.

Li W, Wu J, Tao W, Zhao C, Wang Y*, et al.* (2007). A genetic and bioinformatic analysis of Streptomyces coelicolor genes containing TTA codons, possible targets for regulation by a developmentally significant tRNA. *FEMS Microbiol Lett* **266**: 20-8.

Li YD, Xie ZY, Du YL, Zhou Z, Mao XM*, et al.* (2009b). The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonymous and synonymous sites. *Gene* **436**: 8-11.

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

Lipschultz F and Owens NJP (1996). An assessment of nitrogen fixation as a source of nitrogen to the North Atlantic Ocean. *Biogeochemistry* **35**: 261-274.

Lockless SW and Ranganathan R (1999). Evolutionarily conserved pathways of energetic connectivity in protein families.

Science **286**: 295-9.

Long SR (1989). Rhizobium-legume nodulation: life together in the underground. *Cell* **56**: 203-214.

Long SR (2001). Genes and signals in the Rhizobium-legume symbiosis. *Plant Physiol* **125**: 69-72.

Lowery RG and Ludden PW (1988). Purification and properties of dinitrogenase reductase ADP-ribosyltransferase from the photosynthetic bacterium Rhodospirillum rubrum. *J Biol Chem* **263**: 16714-9.

Lowery RG and Ludden PW (1989). Effect of nucleotides on the activity of dinitrogenase reductase ADP-ribosyltransferase from Rhodospirillum rubrum. *Biochemistry* **28**: 4956-61.

Luscombe NM, Greenbaum D and Gerstein M (2001). What is bioinformatics? An introduction and overview. *Yearb Med Inform* **1**: 83-99.

Ma Y and Ludden PW (2001). Role of the dinitrogenase reductase arginine 101 residue in dinitrogenase reductase ADP-ribosyltransferase binding, NAD binding, and cleavage. *J Bacteriol* **183**: 250-6.

MacLean AM, Finan TM and Sadowsky MJ (2007). Genomes of the symbiotic nitrogen -fixing of Bacteria legumes. *Plant Physiol.* **144**: 615-622.

Mahe F, Markova D, Pasquet R, Misset MT and Ainouche A (2011). Isolation, phylogeny and evolution of the SymRK gene in the legume genus Lupinus L. *Mol Phylogenet Evol* **60**: 49-61.

Mao C, Qiu J, Wang C, Charles TC and Sobral BW (2005). NodMutDB: a database for genes and mutants involved in symbiosis. *Bioinformatics* **21**: 2927-9.

Marcus M and Minc H (1988). Introduction to linear algebra, Dover Publications.

Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F*, et al.* (2006). An experimental metagenome data management and analysis system. *Bioinformatics* **22**: e359-67.

Martin-Galiano AJ, Wells JM and de la Campa AG (2004). Relationship between codon biased genes, microarray expression values and physiological characteristics of Streptococcus pneumoniae. *Microbiology* **150**: 2313-25.

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

Mathesius U, Keijzers G, Natera SH, Weinman JJ, Djordjevic MA*, et al.* (2001). Establishment of a root proteome reference map for the model legume Medicago truncatula using the expressed sequence tag database for peptide mass fingerprinting. *Proteomics* **1**: 1424-40.

Mathura M and Tuli R (1991). Analysis of codon usage in genes for nitrogen fixation from phylogenetically diverse diazotrophs. *J Mol Evol* **32**: 364-373.

Mayer SM, Lawson DM, Gormal CA, Roe SM and Smith BE (1999). New insights into structure-function relationships in nitrogenase: A 1.6 A resolution X-ray crystallographic study of Klebsiella pneumoniae MoFe-protein. *J Mol Biol* **292**: 871-91.

Mazur BJ, Rice D and Haselkorn R (1980). Identification of blue-green algal nitrogen fixation genes by using heterologous DNA hybridization probes. *Proc Natl Acad Sci U S A* **77**: 186-90.

McCammon JA, Gelin BR and Karplus M (1977). Dynamics of folded proteins. *Nature* **267**: 585-90.

Medini D, Donati C, Tettelin H, Masignani V and Rappuoli R (2005). The microbial pan -genome. *Curr Opin Genet Dev* **15**: 589-94.

Meeks JC, Elhai J, Thiel T, Potts M, Larimer F*, et al.* (2001). An overview of the genome of Nostoc punctiforme, a multicellular, symbiotic cyanobacterium. *Photosynth Res* **70**: 85-106.

Merrick MJ, Arcondeguy T and van Heeswijk WC (2000). Regulation of nif gene expression in free-living diazotrophs: Recent advances. In Nitrogen Fixation: From Molecules to crop productivity. F. O. Pedrosa, M. Hungria, and W. E. Newton, Kluwer Academic Publishers**:** 67-72.

Meuwly M and Karplus M (2004). Theoretical investigations on Azotobacter vinelandii ferredoxin I: effects of electron transfer on protein dynamics. *Biophys J* **86**: 1987-2007.

Miflin BJ and Lea PJ (1976). The pathway of nitrogen assimilation in plants. *Phytochemistry* **15**: 873-885.

Milenkovic VM, Brockmann M, Stohr H,

Weber BH and Strauss O (2010). Evolution and functional divergence of the anoctamin family of membrane proteins. *BMC Evol Biol* **10**: 319.

Miyamoto S and Kollman PA (1992). SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J Comput Chem* **13**: 952-962.

Mondal UK, Das B, Ghosh TC, Sen A and Bothra AK (2008). Nucleotide triplet based molecular phylogeny of class I and class II aminoacyl t-RNA synthetase in three domain of life process: bacteria, archaea, and eukarya. *J Biomol Struct Dyn* **26**: 321-8.

Mondal UK, Sen A and Bothra AK (2011). Characterization of Pathogenic Genes through Condensed Matrix Method, Case Study through Bacterial Zeta Toxin. *Int J Genet Eng Biotechnol.* **2**: 109-114.

Morett E and Buck M (1989). In vivo studies on the interaction of RNA polymerase-sigma 54 with the Klebsiella pneumoniae and Rhizobium meliloti nifH promoters. *J Mol Biol* **210**: 65-77.

Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE*, et al.* (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**: 1639 -1662.

Mulligan ME and Haselkorn R (1989). Nitrogen fixation (nif) genes of the cyanobacterium Anabaena species strain PCC 7120. The nifB-fdxN-nifS-nifU operon. *J Biol Chem* **264**: 19200-7.

Murzin AG, Brenner SE, Hubbard T and Chothia C (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**: 536-40.

Mylona P, Pawlowski K and Bisseling T (1995). Symbiotic Nitrogen Fixation. *Plant Cell* **7**: 869-885.

Myrold DD and Huss-Danell K (2003). Alder and lupine enhance nitrogen cycling in a degraded forest soil in Northern Sweden. *Plant and soil* **254**: 47-56.

Nadler SA (1995). Advantages and disadvantages of molecular phylogenetics: a case study of ascaridoid nematodes. *J Nematol* **27**: 423-32.

Natera SHA, Guerreiro N and Djordjevic MA (2000). Proteome analysis of differentially displayed proteins as a tool for the investigation of symbiosis. *Mol Plant Microbe Interact* **13**: 995-1009.

Nei M and Gojobori T (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**: 418-26.

Nielsen R and Yang Z (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**: 929-36.

Normand P and Bousquet J (1989). Phylogeny of nitrogenase sequences in Frankia and other nitrogen-fixing microorganisms. *J Mol Evol* **29**: 436-47.

Normand P, Gouy M, Cournoyer B and Simonet P (1992). Nucleotide sequence of nifD from Frankia alni strain ArI3: phylogenetic inferences. *Mol Biol Evol* **9**: 495-506.

Normand P, Lapierre P, Tisa LS, Gogarten JP, Alloisio N*, et al.* (2007). Genome characteristics of facultatively symbiotic Frankia sp. strains reflect host range and host plant biogeography. *Genome Res* **17**: 7-15.

Normand P, Queiroux C, Tisa LS, Benson DR, Rouy Z*, et al.* (2007). Exploring the genomes of Frankia. *Physiol Plantarum* **130**: 331-343.

Norrby PO and Brandt P (2001). Deriving force field parameters for coordination complexes. *Coord Chem Rev* **212**: 79-109.

O'Carroll IP and Dos Santos PC (2011a). Genomic analysis of nitrogen fixation. *Methods Mol Biol* **766**: 49-65.

O'Carroll IP and Dos Santos PC (2011b). Genomic Analysis of Nitrogen Fixation. *Methods Mol Biol.* **766**: 49-65.

Okayama T, Tamura T, Gojobori T, Tateno Y, Ikeo K*, et al.* (1998). Formal design and implementation of an improved DDBJ DNA database with a new schema and object-oriented library. *Bioinformatics* **14**: 472-8.

Oliveira LR, Marcelino FC, Barcellos FG, Rodrigues EP, Megias M*, et al.* (2010). The nodC, nodG, and glgX genes of Rhizobium tropici strain PRF 81. *Funct Integr Genomics* **10**: 425-31.

Parker MA, Lafay B, Burdon JJ and van Berkum P (2002). Conflicting phylogeographic patterns in rRNA and

nifD indicate regionally restricted gene transfer in Bradyrhizobium. *Microbiology* **148**: 2557-65.

Paul EA and Clark FE (1996). Soil Microbiology and Biochemistry. San Diego, CA, Academic Press.

Paul W and Merrick M (1989). The roles of the nifW, nifZ and nifM genes of Klebsiella pneumoniae in nitrogenase biosynthesis. *Eur J Biochem* **178**: 675-82.

Pawlowski K (2009). Prokaryotic symbionts in plants. Berlin; Heidelberg, Springer.

Pazos F and Valencia A (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng* **14**: 609-14.

Pearson WR and Lipman DJ (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85**: 2444-8.

Peden JF (1999). Analysis of codon usage, University of Nottingham. **PhD**.

Pedrosa FO, Hungria M, Yates G and Newton WE (2000). Nitrogen Fixation: From Molecules to Crop Productivity, Springer.

Peoples MB and Craswell ET (1992). Biological nitrogen fixation: Investments, expectations and actual contributions to agriculture. *Plant and Soil* **141**: 13-39.

Peoples MB, Herridge DF and Ladha JK (1995). Biological nitrogen fixation: An efficient source of nitrogen for sustainable agricultural production? *Management of Biological Nitrogen Fixation for the Development of More Productive and Sustainable Agricultural Systems*3-28.

Peralta H, Guerrero G, Aguilar A and Mora J (2011). Sequence variability of Rhizobiales orthologs and relationship with physico-chemical characteristics of proteins. *Biol Direct* **6**: 48.

Perez-Iratxeta C, Andrade-Navarro MA and Wren JD (2007). Evolving research trends in bioinformatics. *Brief Bioinform* **8**: 88-95.

Perret X, Freiberg C, Rosenthal A, Broughton WJ and Fellay R (1999). High-resolution transcriptional analysis of the symbiotic plasmid of Rhizobium sp. NGR234. *Mol Microbiol* **32**: 415-25.

Perret X, Staehelin C and Broughton WJ (2000). Molecular basis of symbiotic promiscuity. *Microbiol Mol Biol Rev* **64**: 180-201.

Perriere G and Thioulouse J (2002). Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res* **30**: 4548-55.

Peters JW, Fisher K and Dean DR (1995). Nitrogenase structure and function: a biochemical-genetic perspective. *Annu Rev Microbiol* **49**: 335-66.

Peters JW, Stowell MH, Soltis SM, Finnegan MG, Johnson MK, *et al.* (1997). Redox-dependent structural changes in the nitrogenase P-cluster. *Biochemistry* **36**: 1181-7.

Pope MR, Murrell SA and Ludden PW (1985). Covalent modification of the iron protein of nitrogenase from Rhodospirillum rubrum by adenosine diphosphoribosylation of a specific arginine residue. *Proc Natl Acad Sci U S A* **82**: 3173-7.

Postgate JR (1982). The fundamentals of nitrogen fixation. Cambridge, Cambridge University Press.

Postgate JR (1998). Nitrogen fixation. Cambridge, U.K. ; New York, NY, USA, Cambridge University Press.

Postgate JR and Eady RR (1988). The evolution of biological nitrogen fixation. In Bother,H. et al. (eds), Nitrogen Fixation: Hundred Years After. Proceedings of the 7th international congress on nitrogen fixation. Gustav Fischer, Stuttgart, pp 31-40

Puhler A, Arlat M, Becker A, Göttfert M , Morrissey JP, *et al.* (2004). What can bacterial genome research teach us about bacteria-plant interactions? *Curr Opin Plant Biol* **7**: 137-147.

Qi J, Wang B and Hao BI (2004). Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J Mol Evol* **58**: 1-11.

Qian J, Kwon SW and Parker MA (2003). rRNA and nifD phylogeny of Bradyrhizobium from sites across the Pacific Basin. *FEMS Microbiol Lett* **219**: 159-65.

Quackenbush J, Liang F, Holt I, Pertea G and Upton J (2000). The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res* **28**: 141-145.

Raes J and Van de Peer Y (2003). Gene duplication, the evolution of novel gene functions, and detecting functional

divergence of duplicates in silico. *Appl Bioinformatics* **2**: 91-101.

Rai AN, Bergman B and Rasmussen U (2002). Cyanobacteria in symbiosis. Dordrecht; Boston, Kluwer Academic.

Ramseier TM and Gottfert M (1991). Codon usage and G+C content in Bradyrhizobium japonicum genes are not uniform. *Arch Microbiol* **156**: 270-6.

Randic M, Guo X and Basak SC (2001). On the characterization of DNA primary sequences by triplet of nucleic acid bases. *J Chem Inf Comput Sci* **41**: 619-26.

Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, *et al.* (2008). The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *J Bacteriol* **190**: 6881-93.

Raymond J, Siefert JL, Staples CR and Blankenship RE (2004). The natural history of nitrogen fixation. *Mol Biol Evol* **21**: 541-54.

Rees DC (2002). Great metalloclusters in enzymology. *Annu Rev Biochem* **71**: 221-46.

Rees DC and Howard JB (2000). Nitrogenase: standing at the crossroads. *Curr Opin Chem Biol* **4**: 559-66.

Rees DC, Akif Tezcan F, Haynes CA, Walton MY, Andrade S, *et al.* (2005). Structural basis of biological nitrogen fixation. *Philos Transact A Math Phys Eng Sci* **363**: 971-84; discussion 1035-40.

Ribbe M, Gadkari D and Meyer O (1997). N2 Fixationby Streptomyces thermoautotrophicus Involves a Molybdenum-Dinitrogenase and a Manganese-Superoxide Oxidoreductase That Couple N2Reduction to the Oxidation of Superoxide Produced from O2by a Molybdenum-CO Dehydrogenase. *J Biol Chem* **272**: 26627-26633.

Rice D, Mazur BJ and Haselkorn R (1982). Isolation and physical mapping of nitrogen fixation genes from the cyanobacterium Anabaena 7120. *J Biol Chem* **257**: 13157-63.

Rijkenberg MJ, Langlois RJ, Mills MM, Patey MD, Hill PG, *et al.* (2011). Environmental forcing of nitrogen fixation in the eastern tropical and sub-tropical North Atlantic Ocean. *PLoS One* **6**: e28989.

Roberts GP, MacNeil T, MacNeil D and Brill WJ (1978). Regulation and characterization of protein products coded by the nif (nitrogen fixation) genes of Klebsiella pneumoniae. *J Bacteriol* **136**: 267-79.

Robson RL, Eady RR, Richardson TH, Miller RW, Hawkins M, *et al.* (1986). The alternative nitrogenase of Azotobacter chroococcum is a vanadium enzyme. *Nature* **322**: 388-390.

Rosswall T (1976). The internal nitrogen cycle between microorganisms, vegetation and soil. *Ecological Bulletins*157-167.

Rubio LM and Ludden PW (2005). Maturation of nitrogenase: a biochemical puzzle. *J Bacteriol* **187**: 405-14.

Rubio LM and Ludden PW (2008). Biosynthesis of the iron-molybdenum cofactor of nitrogenase. *Annu Rev Microbiol* **62**: 93-111.

Rubio LM, Ludden PW and Leigh GJ (2002). The gene products of the *nif* regulon, p. 101-136. *In* G. J. Leigh (ed.),Nitrogen fixation at the millenium. Elsevier, Amsterdam, The Netherlands.

Russ W and Ranganathan R (2010). Evolution-guided engineering of serine protease specificity. *Neuro Informatics 2010*123.

Russell RB and Sternberg MJ (1996). A novel binding site in catalase is suggested by structural similarity to the calycin superfamily. *Protein Eng* **9**: 107-11.

Saitou N and Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406-25.

Sali A (1995). Modeling mutations and homologous proteins. *Curr Opin Biotechnol* **6**: 437-51.

Schindelin H, Kisker C, Schlessman JL, Howard JB and Rees DC (1997). Structure of ADP x AIF4(-)-stabilized nitrogenase complex and its implications for signal transduction. *Nature* **387**: 370-6.

Schlegel HG and Zaborosch C (2003). General microbiology. Cambridge, Cambridge University Press.

Schlessman JL, Woo D, Joshua-Tor L, Howard JB and Rees DC (1998). Conformational variability in structures of the nitrogenase iron proteins from Azotobacter vinelandii and Clostridium pasteurianum. *J Mol Biol* **280**: 669-85.

Schmid B, Einsle O, Chiu HJ, Willing A,

Yoshida M, *et al.* (2002). Biochemical and structural characterization of the cross-linked complex of nitrogenase: comparison to the ADP-AlF4(-)-stabilized structure. *Biochemistry* **41**: 15557-65.

Schultze M and Kondorosi A (1998). Regulation of symbiotic root nodule development. *Annu Rev Genet* **32**: 33-57.

Scott WRP, Hunenberger PH, Tironi IG, Mark AE, Billeter SR, *et al.* (1999). The GROMOS biomolecular simulation program package. *J Phys Chem A* **103**: 3596-3607.

Searls DB (2000). Bioinformatics tools for whole genomes. *Annu Rev Genomics Hum Genet* **1**: 251-79.

Sen A, Sur S, Bothra AK, Benson DR, Normand P, *et al.* (2008). The implication of life style on codon usage patterns and predicted highly expressed genes for three Frankia genomes. *Anton Van Leeuwen* **93**: 335-46.

Sen A, Sur S, Tisa L, Bothra A, Thakur S, *et al.* (2010). Homology modelling of the Frankia nitrogenase iron protein. *Symbiosis* **50**: 37-44.

Sen A, Thakur S, Bothra AK, Sur S and Tisa LS (2012). Identification of TTA codon containing genes in Frankia and exploration of the role of tRNA in regulating these genes. *Arch Microbiol* **194**: 35-45.

Sen S and Peters JW (2006). The thermal adaptation of the nitrogenase Fe protein from thermophilic Methanobacter thermoautotrophicus. *Proteins* **62**: 450-60.

Sen S, Igarashi R, Smith A, Johnson MK, Seefeldt LC, *et al.* (2004). A conformational mimic of the MgATP-bound "on state" of the nitrogenase iron protein. *Biochemistry* **43**: 1787-97.

Sharp PM and Li WH (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**: 1281-95.

Sharp PM and Li WH (1987). The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* **4**: 222-30.

Sharp PM, Tuohy TM and Mosurski KR (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* **14**: 5125-43.

Silver WS and Postgate JR (1973). Evolution of asymbiotic nitrogen fixation. *J Theor Biol* **40**: 1-10.

Sims GE, Jun SR, Wu GA and Kim SH (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci U S A* **106**: 2677-82.

Smock RG, Rivoire O, Russ WP, Swain JF, Leibler S, *et al.* (2010). An interdomain sector mediating allostery in Hsp70 molecular chaperones. *Mol Syst Biol* **6**: 414.

Snipen L and Ussery DW (2010). Standard operating procedure for computing pangenome trees. *Stand Genomic Sci* **2**: 135-41.

Snipen L, Almoy T and Ussery DW (2009). Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics* **10**: 385.

Steinway SN, Dannenfelser R, Laucius CD, Hayes JE and Nayak S (2010). JCoDA: a tool for detecting evolutionary selection. *BMC Bioinformatics* **11**: 284.

Stewart WD (1969). Biological and ecological aspects of nitrogen fixation by free-living micro-organisms. *Proc R Soc Lond B Biol Sci* **172**: 367-88.

Strop P, Takahara PM, Chiu H, Angove HC, Burgess BK, *et al.* (2001). Crystal structure of the all-ferrous [4Fe-4S]0 form of the nitrogenase iron protein from Azotobacter vinelandii. *Biochemistry* **40**: 651-6.

Suel GM, Lockless SW, Wall MA, and Ranganathan R (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct Biol* **10**:59–69

Suhre K and Sanejouand YH (2004). ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res* **32**: W610-W614.

Sundaram S, Tripathi A and Gupta V (2010). Structure prediction and molecular simulation of gases diffusion pathways in hydrogenase. *Bioinformation* **5**: 177-83.

Sur S, A. Pal, A.K. Bothra and A. Sen (2005). Moderate codon bias attributed to translational selection in nitrogen fixing genes of Bradyrhizobium japonicum USDA110. *Bioinformatics Ind.* **3**: 59-64.

Sur S, Bhattacharya M, Bothra AK, Tisa LS and Sen A (2008). Bioinformatic analysis of codon usage patterns in a free-living diazotroph, Azotobacter vinelandii. *Biotechnology* **7**: 242-249.

Sur S, Bothra AK, Ghosh TC and Sen A (2010). Investigation of the molecular evolution of nitrogen fixation using nucleotide triplet based condensed matrix method. *Int. J. Integrative Biol.* **10**: 29-65.

Suyama M, Torrents D and Bork P (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**: W609-12.

Tama F and Sanejouand YH (2001). Conformational change of proteins arising from normal mode calculations. *Protein Eng* **14**: 1-6.

Tamura K, Dudley J, Nei M and Kumar S (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**: 1596-9.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B*, et al.* (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**: 41.

Tatusov RL, Galperin MY, Natale DA and Koonin EV (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**: 33-6.

Tennessen JA (2008). Positive selection drives a correlation between non-synonymous/ synonymous divergence and functional divergence. *Bioinformatics* **24**: 1421-5.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D*, et al.* (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**: 13950-5.

Tettelin H, Riley D, Cattuto C and Medini D (2008). Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* **11**: 472-7.

Thakur S, Bothra AK and Sen A (2012). In silico studies of NifH protein structure and its post-translational modification in Bradyrhizobium sp. ors278. *Int J Pharm Bio Sc* **3**: B22 - B32.

Thompson JD, Higgins DG and Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673-80.

Thorneley RNF and Lowe DJ (1985). Kinetics and mechanism of the nitrogenase enzyme system. *Molybdenum enzymes*221-284.

Torrey JG (1978). Nitrogen Fixation by Actinomycete-Nodulated Angiosperms. *BioScience* **28**: 586-592.

Torrey JG and Tjepkema JD (1979). Symbiotic Nitrogen Fixation in Actinomycete-Nodulated Plants. Program, The University of Chicago Press. **140**.

Uchiumi T, Ohwada T, Itakura M, Mitsui H, Nukui N*, et al.* (2004). Expression islands clustered on the symbiosis island of the Mesorhizobium loti genome. *J Bacteriol* **186**: 2439-48.

Ussery DW, Borini S and Wassenaar T (2009). Computing for comparative genomics: bioinformatics for microbiologists. *London: Spring-verlag*.

Vitousek P, Cassman K, Cleveland C, Crews T, Field C*, et al.* (2002). Towards an ecological understanding of biological nitrogen fixation. *Biogeochemistry* **57-58**: 1-45.

Vitousek PM, Hattenschwiler S, Olander L and Allison S (2002). Nitrogen and nature. *Ambio* **31**: 97-101.

Von Bulow JF and Dobereiner J (1975). Potential for nitrogen fixation in maize genotypes in Brazil. *Proc Natl Acad Sci U S A* **72**: 2389-93.

Walker SA, Viprey V and Downie JA (2000). Dissection of nodulation signaling using pea mutants defective for calcium spiking induced by nod factors and chitin oligomers. *Proc Natl Acad Sci U S A* **97**: 13413-8.

Wang R, Prince JT and Marcotte EM (2005). Mass spectrometry of the M. smegmatis proteome: protein expression levels correlate with function, operons, and codon bias. *Genome Res* **15**: 1118-26.

Wani SP, Rupela OP and Lee KK (1995). Sustainable agriculture in the semi-arid tropics through biological nitrogen fixation in grain legumes. *Plant and soil* **174**: 29-49.

Watanabe I and Liu CC (1992). Improving nitrogen-fixing systems and integrating them into sustainable rice farming. *Plant and Soil* **141**: 57-67.

Watson JD and Milner-White EJ (2002). The conformations of polypeptide chains where the main-chain parts of successive residues are enantiomeric. *J Mol Biol* **315**: 183-91.

Wheeler CT and Miller IM (1990). Current potential uses of actinorhizal plants in Europe. The biological of Frankia and actinorhizal plants. R. C. Schwintzer and J. D. Tjepkema. San Diego, Academic press Inc**:** 365-389.

Wiederstein M and Sippl MJ (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* **35**: W407-10.

Worth CL, Preissner R and Blundell TL (2011). SDM--a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res* **39**: W215-22.

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

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

Yang GP, Debelle F, Savagnac A, Ferro M, Schiltz O*, et al.* (1999). Structure of the Mesorhizobium huakuii and Rhizobium galegae Nod factors: a cluster of phylogenetically related legumes are nodulated by rhizobia producing Nod factors with alpha,beta-unsaturated N-acyl substitutions. *Mol Microbiol* **34**: 227-37.

Yang Z (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**: 555-6.

Yang Z (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586-91.

Yang Z, Nielsen R, Goldman N and Pedersen AM (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**: 431-49.

Yates RJ, Howieson JG, Reeve WG and O'Hara GW (2011). A re-appraisal of the biology and terminology describing rhizobial strain success in nodule occupancy of legumes in agriculture. *Plant and soil* **348**: 255-267.

Yi J (2009). The Medicago truncatula genome and analysis of nodule-specific genes, The University of Oklahoma**:** 141.

Young J (2005). The phylogeny and evolution of nitrogenases. *Genomes and genomics of nitrogen-fixing organisms*221-241.

Young JP, Crossman LC, Johnston AW, Thomson NR, Ghazoui ZF*, et al.* (2006). The genome of Rhizobium leguminosarum has recognizable core and accessory components. *Genome Biol* **7**: R34.

Young JPW (1992). Phylogenetic classification of nitrogen-fixing organisms. Biological Nitrogen Fixation. G. Stacey, R. H. Burris and H. J. Evans. New York, Chapman & Hall**:** 43-86.

Young ND, Debelle F, Oldroyd GE, Geurts R, Cannon SB*, et al.* (2012). The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature* **480**: 520-4.

Zehr JP, Harris D, Dominic B and Salerno J (1997). Structural analysis of the Trichodesmium nitrogenase iron protein: implications for aerobic nitrogen fixation activity. *FEMS microbiol lett* **153**: 303-309.

Zehr JP, Mellon MT and Hiorns WD (1997b). Phylogeny of cyanobacterial nifH genes: evolutionary implications and potential applications to natural assemblages. *Microbiology* **143 (Pt 4)**: 1443-50.

Zehr JP, Jenkins BD, Short SM and Steward GF (2003). Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Environ Microbiol* **5**: 539-54.

Zehr JP (2011). Nitrogen fixation by marine cyanobacteria. *Trends Microbiol* **19**: 162-73.

Zhao H, Li M, Fang K, Chen W and Wang J (2012). In Silico Insights into the Symbiotic Nitrogen Fixation in Sinorhizobium meliloti via Metabolic Reconstruction. *PLoS One* **7**: e31287.

Zheng L, White RH and Dean DR (1997). Purification of the Azotobacter vinelandii nifV-encoded homocitrate synthase. *J Bacteriol* **179**: 5963-6.

Zheng Y, Xu D and Gu X (2007). Functional divergence after gene duplication and sequence-structure relationship: a case study of G-protein alpha subunits. *J Exp Zool B Mol Dev Evol* **308**: 85-96.

Ziegel ER (1993). Correspondence Analysis Handbook. *Technometrics* **35**: 103-103.

Index

Appendix-A

Thesis related publication till Dec, 2012

Published

- **Thakur S**, AK Bothra and A Sen (2012). Insights into the sequence and structural features of nitrogenase protein - the *in silico* approach. In ed. A. Sen *Biology of useful plants and microbes*. Narosa Pubshling House Pvt Ltd. (In Press)
- **Thakur S**, AK Bothra and A Sen (2012). *In silico* studies of NifH protein structure and its post-translational modification in Bradyrhizobium sp. ORS278. *Int. J. Pharma and BioSciences* **3**: B22-B32.
- Sen A, **S Thakur**, AK Bothra, S Sur and LS Tisa (2012). Characterization of TTA codon containing genes in Frankia and exploration of the role of tRNA in regulating these genes. *Archives of Microbiology* **194**:35-45.
- **Thakur S**, Z Patra, M Biswal, AK Bothra and A Sen (2011). Characterization of proteins with signal peptides in some actinomycetes- a bioinformatics approach. *Int. J. Integrative Biol.* **11**: 105-109
- A Sen, S Sur, LS Tisa, AK Bothra, **S Thakur** and UK Mondal (2009). Homology modeling of the *Frankia* nitrogenase iron protein. *Symbiosis* 50:37-44

Submitted**:**

- **Thakur** S, P Normand, V Daubin, L Tisa, and A Sen (2012). Contrasted evolutionary constraints on secreted and non-secreted proteomes of *Frankia* spp. (Manuscript communicated to *BMC Genomics*)
- **Thakur S**, AK Bothra and A Sen (2012). Functional divergence outlines the evolution of novel protein function in NifH/BchL protein family. (Manuscript communicated to *Int J Bioinform Res Appl*)
- **Thakur S**, AK Bothra and A Sen (2012) Exploring the genomes of symbiotic diazotrophs with relevance to biological nitrogen fixation (Manuscript communicated for book chapter (springer publication))

Appendix-B

Software used in the present study

Name	Executable	Description
CodonW	Windows	Program for codon and amino acid usage
Modeller	Windows	Standalone program for homology modeling
Pymol	Windows	Molecular visualization software written in python
Rasmol	Windows	Molecular visualization software
GROMACS	Linux	Molecular dynamics package
Amber	Linux	Programs for molecular dynamics simulations
DSSP	Linux	Program for secondary structure alignment
Gaussian03	Windows	Quantum chemistry software package
Gamess	Linux	*Ab* initio quantum chemistry package
XLSTAT	Windows	Statistical and data analysis software package
PAML	Linux	Phylogenetic analysis using maximum likelihood
JCoDA	Windows	Tool for detecting evolutionary selection
CMG-biotools	Linux	Stand alone OS for comparative microbial genomics
MEGA	Windows	Tool for sequence alignment and phylogeny
ClustalW	Windows	Multiple sequence alignment program
Bioedit	Windows	Biological sequence alignment editor
AutoDock 4.0	Linux/Windows	Suite of automated docking tools
MATLAB	Windows	Numerical computing environment
SPSS	Windows	Software package used for statistical analysis
Phylip	Windows	Phylogeny Inference Package computer program
Phylodraw	Windows	Phylogenetic tree drawing system
FigTree	Windows	Graphical viewer of phylogenetic trees
Diverge	Windows	Program for detection of functional divergence
PhyML	Windows	Estimating ML phylogenies from alignments
Swiss-PDB Viewer	Windows	Environment for comparative protein modeling
RNAmmer	Windows	Program for rRNA gene prediction
BLAST	Windows/linux	Algorithm for local similarity between sequences
R-Package	Windows/linux	Functional language for statistical analysis
GraphPad	Windows	Software for plotting graphs
DNA-plotter	Windows	Software for Genome visualization

Appendix-C

Web server used in present study

Name	Web Address	Description
JGI-IMG	www.img.jgi.doe.gov	Integrated Microbial Genomes system
NCBI	www.ncbi.nlm.nih.gov/	For molecular biology information
PDB	www.rcsb.org/	Repository of 3D protein structures
ProFunc	http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/	Identifies biochemical function of protein from its 3D structure
VERIFY3D	http://nihserver.mbi.ucla.edu/Verify_3D/	Aids in the refinement of 3D structures.
ERRAT	http://nihserver.mbi.ucla.edu/ERRATv2/	Protein structure verification algorithm
ProSA	https://prosa.services.came.sbg.ac.at/prosa.php	Refinement and validation of crystallographic structures
DALI	http://ekhidna.biocenter.helsinki.fi/dali_server/	Comparing protein structures in 3D
CASTp server	http://sts.bioengr.uic.edu/castp/	Identification of protein pockets
PAL2NAL	http://www.bork.embl.de/pal2nal/	Calculation of dN/dS
CAI Calculator2	http://userpages.umbc.edu/~wug1/codon/cai/cais.php	Calculation of codon adaptation index
ASA view	http://gibk26.bio.kyutech.ac.jp/jouhou/shandar/netasa/asaview/	Provides graphical representation of solvent accessibility
HNN	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_hnn.html	Secondary structure prediction method
SDM	http://mordred.bioc.cam.ac.uk/~sdm/sdm.php	Calculates stability score for mutant proteins
WEBnm@	http://www.bioinfo.no/tools/normalmodes	Web application for NMA of proteins
ElNe´mo	http://igs-server.cnrs-mrs.fr/elnemo/index.html	Tool for prediction of protein movements
RESP	http://q4md-forcefieldtools.org/REDS/	Server for charge derivation
SWISS-MODEL repository	http://swissmodel.expasy.org/repository/	Database for 3D homology model

Appendix-D

Scripts utilized in the present study

In-house developed scripts:

Name	Description
KaKsmod.pl	Perl script for determining of pairwise Ka/Ks value by running codeml program
Ortho.pl	Perl script for parsing blast results for ortholog detection
Nif.pl	Perl script for parsing for *nif* genes from genomes
Removeblk.pl	Perl script for removing blank spaces in genome information
Rarecodon.pl	Perl script for estimating the frequency of rare codon
Genebank.pl	Perl script for retrieving gene sequences from genbank in fasta format.
GANAL.cpp and Sau3.cpp*	Turbo C++ scripts for count each and every potential triplet from nucleotide sequences

Scripts developed by others:

Name	Description
freqaa.pl	Perl script used for determination of amino-acid compositions
Tree.pl	Perl script for pangenome tree
Specific_gene.pl	Perl script retrieving pan-genome and core-genome
Roseplot.R	R-script for rose plot construction
Sca.m	Matlab script for SCA calculatiion

Appendix-E

List of Supplementary Material
(provided in the compact disc attached with the thesis)

Supplementary Tables

Table ST1	Mean values of Nc, GC, GC3, CAI, CBI and Fop for the genes in the studied diazotroph
Table ST2	Amino acid compositional matrix for different diazotrophs
Table ST3-ST6	Noe, Nol, Fix, Nif ortholog in Rhizobiales
Table ST7	Orthologs of Nitrogen fixation related genes in *Frankia* genomes
Table ST8-10	Distance matrix for the nifH, nifD & nifK genes

Supplementary Figures

Figure SF 1.1-1.4	Nc/GC3 plots for various diazotrophs
Figure SF 2.1-2.10	Codon usage and Amino acid usage rose plots for symbiotic diazotroph and cyanobacteria
Figure SF 3.1-3.3	Sliding window analysis of pairwise Ka/Ks in *nif*HDK gene
Figure SF 4.1-4.5	Phylogram of nifHDK genes using sequence based method and nucleotide triplet based method
Figure SFHR 4.8	High resolution figure for the blast matrix involving 15 Rhizobiales members