In-Silico Functional Annotation of Unannotated Protein-coding Gene of *Arthrospira platensis* NIES-39 Genome and Investigation of Proteins Involved in Nitrogen Assimilation Pathway

THESIS

Submitted in partial fulfillment

of the requirements for the degree of

DOCTOR OF PHILOSOPHY

By

PARVA KUMAR SHARMA

2011PHXF0407P

Under the Supervision of

Prof. Shibasish Chowdhury

BITS Pilani

Pilani | Dubai | Goa Hyderabad

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

2018

Birla Institute of Technology & Science
Pilani (Rajasthan)

Certificate

This is to certify that the thesis entitled "***In-Silico*** **Functional Annotation of Unannotated Protein-coding Gene of** ***Arthrospira platensis*** **NIES-39 Genome and Investigation of Proteins Involved in Nitrogen Assimilation Pathway**" and submitted by **Parva Kumar Sharma, ID No. 2011PHXF0407P** for award of Ph.D. degree of the Institute embodies original work done by him under my supervision.

Signature of the supervisor : Shibasish Chowdhury

Name (capital letters) : SHIBASISH CHOWDHURY

Designation : Associate Professor

Date : 6/6/19

Acknowledgement

It's been a wonderful and joyful experience throughout my PhD degree be it academic, social or personal. Starting from the very first day of my PhD to the end of the work, a lot of people helped and supported me during my research. I would like to acknowledge all of them for their support and help throughout my research work.

The very first figure comes to my mind is the creator of the universe, whose regular inspiration in terms of thoughts and patience helped me with this journey. Secondly, my Guru Pandit Sri Ram Sharma Acharya, whose life and literature both shaped (and still shaping) my life.

I am thankful to the Vice Chancellor, Directors and Deans of various divisions at Birla Institute of Technology and Science (BITS), Pilani for providing necessary facilities and financial support. Special thanks to Prof. Srinivas Krishnaswamy, Dean, Academic Graduate Studies and Research Division (AGSRD), BITS Pilani, Pilani Campus and Prof. Shibasish Chowdhury, Convener, Departmental Research Committee (DRC), Department of Biological Sciences, BITS Pilani, Pilani Campus for their official support and personal encouragement. I also acknowledge Prof. Jitendra Panwar, Associate Dean, AGSRD and his office staff Mr. Mahipal and Mr. Raghubir for their cooperation and constant assistance for official work.

From the core of my heart, I would like to express my sincere gratitude towards my supervisor Prof. Shibasish Chowdhury, without whom, this thesis would not have been a success. His suggestions, encouragements, support and constructive criticisms helped me in getting most out of my potential. I am also grateful to him for giving me the liberty to carry out my research work independently throughout the tenure. He has truly been a fantastic mentor throughout my research. I am also thankful to my Doctoral Advisory Committee (DAC) members Dr. B. Vani and Dr. Sandhya Mehrotra who were always there to support me and spared their valuable time to proof-read this thesis. Their honest comments and useful suggestions have immensely helped in the enrichment of this thesis.

I would like to acknowledge all the faculty members of the Department of Biological Sciences BITS, Pilani for their moral support and encouragement throughout my research work. I express my thanks to the office staff members of the Department of Biological Sciences for their help and cooperation.

I would like to thank all my seniors at BITS Pilani for their valuable advice and who always inspired me at my stay at BITS Pilani. I would also like to thank my friend and co-scholar

Mr. Rajnish Singh for his time to time support and encouragement. I would like to thank my lab mates Ms. Divya and Miss Harshita for their support and help in research as well as in other academic works. I would like to thank all the research scholars of the Department of Biological Sciences for their moral support and help during my PhD work. I would like to acknowledge my friend Vishnu Vardhan for editing this thesis.

None of this would have been possible without the blessing and support of my family. I express my special indebtedness towards my parents Mr. P. K. Sharma and Mrs. Priyamvada Sharma for their love, support, trust, encouragement and tremendous patience towards me. My due thanks to my brother Puneet, who constantly kept me checked. Special thanks is reserved for my wife Sonu who always stood beside me in all the ups and downs of life and whose constant support and encouragement kept pushing me in this journey. Last but not the least, my son Shreshth who had no idea what I was doing but whose innocent face always kept inspiring me and is a source of eternal happiness.

I convey my obligation to UGC, New Delhi and BITS, Pilani for their valuable financial support during my research tenure.

Lastly, I wish to apologise if I forgot to acknowledge anyone who had helped me in any way.

Parva Kumar Sharma

Abstract

Arthrospira platensis NIES-39 is non-nitrogen fixing photosynthetic, highly alkalophilic prokaryotic cyanobacterium with high protein content (~65%) which is used as a protein supplement in the human diet. However, about 22% of protein-coding genes of this cyanobacterium are functionally unannotated (hypothetical proteins). These unannotated proteins of this species could hold the vital information regarding its unique characteristic features. Several pathways like mRNA degradation, tRNA synthetase and the nitrogen assimilation pathway have been known to contribute towards the high protein content of a cell. Out of these, the nitrogen assimilation pathway helps in the incorporation of nitrogen into various cellular molecules like amino acids and DNA. However, how can these enzymes in *Arthrospira platensis* NIES-39 play a differential role in the nitrogen assimilation pathway is still not known.

In the present study, we could annotate 526 hypothetical proteins of *Arthrospira platensis* NIES-39 including many functionally important proteins like ABC transporters, transcriptional regulators, restriction endonucleases, metal ion binding, hydrolyzing enzymes, oxidoreductases and helicases. Some of these annotated proteins are known to involve in stress management and protein production pathways.

Sequence, structural and evolutionary analysis of nitrogen assimilatory enzymes of *Arthrospira platensis* NIES-39 was carried out to understand the role of these enzymes on some of the characteristics features of this organism. Sequence analyses have identified conserved patterns in the domains of all the four enzymes. A C-terminal motif was identified in NR of *Arthrospira platensis* NIES-39. Some key residue positions were also identified which could be associated with the final protein content. In NR of *Arthrospira platensis* NIES-39, position 394 was found that could be responsible for its differential functioning in nitrogen assimilation. In NiR, it was found that the position 408 switches the enzyme from low to high affinity. *Arthrospira platensis* NIES-39 has low-affinity NiR. In case of GOGAT of *Arthrospira platensis* NIES-39, an insertion was detected in the GATase domain. Among all the enzymes of nitrogen assimilation, GS was found to be highly conserved. Horizontal gene transfer and speciation events were also detected through evolutionary studies.

This analysis will also help us to understand the unique features of nitrogen assimilatory enzymes of *Arthrospira platensis* NIES-39 which could be related to unique features of this cyanobacterium.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2-OG	2-Oxo Gluterate
4Fe-4S	Iron Sulphur Cluster
ABC	ATP Binding Cassette
ADP	Adenosine DiPhosphate
ALS	Amyotrophic Lateral Sclerosis
ATP	Adenosine Tri Phosphate
BLASTN	Basic Local Alignment Search Tool Nucleotide
BLASTP	Basic Local Alignment Search Tool Protein
BMAA	Beta-MethylAmino-L-Alanine
BNF	Biological Nitrogen Fixation
cAMP	Cyclic Adenosine Mono Phosphate
CDC	Centers for Disease Control and prevention
CDD	Conserved Domain Database
cDNA	Complimentary DNA
CELSS	Controlled Ecological Life Support System
COG	Clusters of Orthologous Groups
DMSOR	Dimethylsulfoxide Reductase
DNA	Deoxy riboNucleic Acids
E value	Expect Value
EBD	Effector Binding Domain
EMBOSS	European Molecular Biology Open Software Suite
Fd	Ferrodoxin
Fdh	Formate Dehydrogenase
FM	Fisch-Margoliash
GAT	Glutamine AmidoTransferase
GOGAT	Glutamate 2-OxoGluterate AminoTransferase
GROMACS	GROningen MAchine for Chemical Simulations
GS	Glutamine Synthetase
gyrB	Gyrase B
HAB	Harmful Algal Bloom
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
INDEL	Insertion or Deletion
ITS	Internal Transcribed Spacer
JTT	Jones Taylor Thornton
MEGA	Molecular Evolutionary Genetics Analysis
MEME	Multiple Expectation Minimization for motif Elicitation
MFS	Major Facilitator Superfamily
MGD	Molybdopterin Guanine Dinucleotide
ML	Maximum Likelihood
MLST	Multi Locus Sequence Typing
MopB	Molybdopterin-Binding

MP	Maximum Parsimony
mRNA	messenger RNA
MSA	Multiple Sequence Alignment
NADH	Nicotinamide Adenine Dinucleotide
NADPH	Nicotinamide Adenine Dinucleotide Phosphate
NAGK	*N*-Acetyl-L-Glutamate Kinase
NAP	Periplasmic NR
NAR	Respiratory NR
NAS	Assimilatory NR
NBD	Nucleotide Binding Domain
NCBI	National Center For Biotechnology Information
N-DOPE	Normalized Discrete Optimized Protein Energy
NIES	National Institute for Environmental Studies
nifH	Nitrogen Fixation H
NiR	Nitrite Reductase
NJ	Neighbor Joining
NNP	Nitrate Nitrite Porter
NR	Nitrate Reductase
NRT	Nitrate Transporters
NtcA	Nitrogen Control A
ompA	Outer Membrane Protein A
PDB	Protein Data Bank
PFAM	Protein Family
PROCHECK	Protein Structure CHECK
PSA	Pairwise Sequence Alignment
PSI-BLAST	Position Specific Iterative - Basic Local Alignment Search Tool
PSSM	Position Specific Scoring Matrix
PTR	Peptide Transporter
rpoB	RNA Polymerase B
rpoC1	RNA Polymerase C1
rRNA	ribosal RiboNucleic Acid
SBP	Substrate Binding Protein
SMART	Simple Modular Architecture Research Tool
snRNA	small nuclear RNA
TIGR	The Institute for Genomic Research
tRNA	trasnsfer RNA
UPGMA	Unweighted Pair Group Method with Arithmatic mean
UTR	UnTrasnlated Region
VMD	Visual Molecular Dynamics
WHO	World Health Organization

Chapter I

Introduction

1.1 Cyanobacteria

Cyanobacteria are one of the oldest groups of organisms found on earth (Schopf and Packer 1987). Initially, they were known as "blue-green algae" due to their resemblance with the eukaryotic green algae. However, later they were identified as prokaryotic organisms that can perform photosynthesis. The blue colour of cyanobacteria is due to the pigment phycocyanin. These ancient organisms are found in almost all the habitats on Earth (Henson et al. 2004) and are considered as one of the most important groups of photoautotrophic bacteria, which have a significant role in natural carbon and nitrogen cycle (Zhang et al. 2018). Cyanobacteria fall in the bacterial division but are very important from an evolutionary point of view as they were present at that point of time in history when environmental and molecular changes were abounding. Consequently, the evolutionary aspects have been trapped into their sequence as well as structural features. These are the only prokaryotes that can perform photosynthesis through flattened sacs called thylakoids and can produce oxygen (Hamilton et al. 2016). Due to their oxygen generating capacity during photosynthesis, cyanobacteria are credited for the oxygenation event which converted the primitive earth's environment from reductive to oxidative (Schopf 2014). Cyanobacteria are also attributed by the presence of chloroplast in the photosynthetic eukaryotes through the process of Endosymbiosis (Ponce-Toledo et al. 2017). Their unique cellular structure places them between prokaryotes and eukaryotes. Thus, Cyanobacteria have both, the simplicity of a prokaryotic cell and details of cellular machinery like that of eukaryotes making them an ideal system to study.

Cyanobacteria are widely used in many processes in biotechnology (Thajuddin and Subramanian 2005, Abed et al. 2009, Pisciotta et al. 2010, Quintana et al. 2011) (Figure 1.1). These are of commercial importance (Mann and Carr 1992) and are widely used in daily human life (Tiffany 1968). Further, cyanobacteria contribute to the water and soil fertility as a primary producer (Rai 2018) and have high nutritional value.

Though cyanobacteria are useful in many ways as described above, it could be harmful as some cyanobacteria produce toxins, called cyanotoxins. These cyanotoxins are chemically and toxicologically diverse compounds. This diversity also reflects in their targets: they could be hepatotoxins like microcystin, nodularin R (from Nodularia), neurotoxins like anatoxin-a and neosaxitoxin cytotoxins, dermatotoxins and irritant toxins (Wiegand and Pflugmacher 2005). Other toxins are aplysiatoxin and domoic acid. Some common species of cyanobacteria that produce cyanotoxins are *Microcystis, Anabaena, and Planktothrix*.

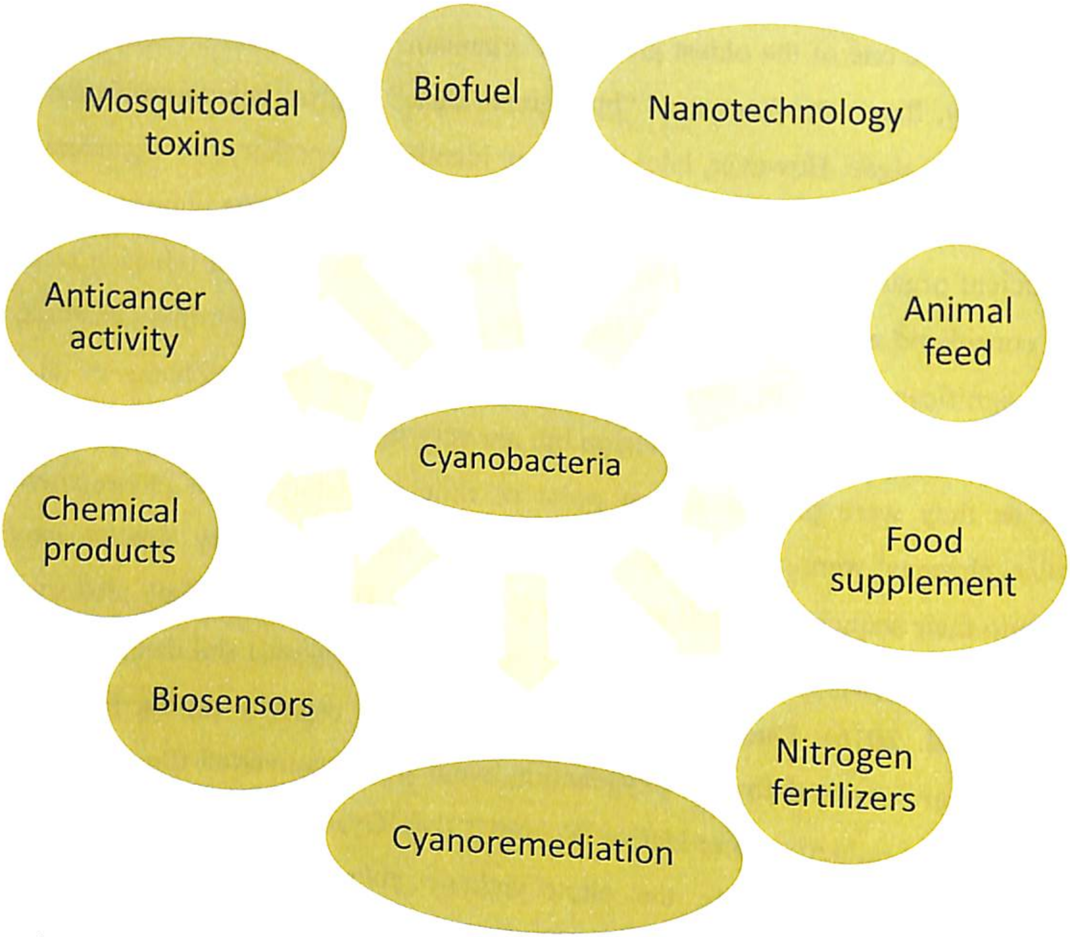


Figure 1.1 Many uses of cyanobacteria. Cyanobacteria have been used in many areas like disease treatment, nanotechnology and food supplement.

Cyanobacteria can form algal blooms under favourable conditions known as harmful algal blooms (HAB) if the cyanobacteria involved in producing toxins (Blaha et al. 2009).

Some human poisoning cases have also been reported (Falconer et al. 1983, Turner et al. 1990, Soong et al. 1992, Elsaadi and Cameron 1993). Several studies suggest that significantly high exposure to some toxin-producing cyanobacterial species can cause some diseases like amyotrophic lateral sclerosis (ALS). A detailed report on the cyanotoxins was published by the World health organisation (WHO) in 1999 (Chorus and Bartram 1999).

Availability of nitrogen is one of the crucial factors for cyanobacterial growth. The cyanobacterial cell contains about 11% of nitrogen of the dry weight (Wolk 1973). Cyanobacteria can utilise various nitrogen-containing compounds as a source of nitrogen. For example, nitrogen-fixing cyanobacteria can directly fix atmospheric nitrogen for their requirements. The non-nitrogen-fixing groups can use either the organic compounds like urea and amino acids or the inorganic compounds like nitrate, nitrite, ammonium and some nitrogen-containing bases as a nitrogen source. Cyanobacteria prefer some particular nitrogen sources over others owing to their easy assimilation inside the cell. This process is known as nitrogen control (Flores and Herrero 2005). Nitrogen acquisition in cyanobacteria consumes photosynthetically generated ATP and reducing power (ferredoxin) (Flores and Herrero 2005).

Cyanobacterial classification had always been under debate. Due to the extensive geographical presence of cyanobacteria, it is challenging to draw the right phylogenetic relationship between them (Dvorak et al. 2015). Traditionally, morphological features are used to infer a phylogenetic relationship among cyanobacteria (Rippka 1988). But nowadays, 16s rRNA gene sequences are the top choice for inferring a phylogenetic relationship (Woese 1987), but some studies also suggest the use of polyphasic approach where multiple genes are used to draw the phylogenetic relationship (Komarek et al. 2014). Multi-Locus Sequence Typing (MLST) like Internal Transcribed Spacer (ITS) and some housekeeping genes (gyrB, rpoC1 and rpoB) are used for phylogenetic analysis (Gaget et al. 2015). Important genes like nifH are also used (Singh et al. 2013) to classify cyanobacteria. Some species trees are also made by concatenating several gene sequences (Gadagkar et al. 2005).

Three main schemes for cyanobacterial classification are:

1. Taxonomic scheme according to Bergey's Manual of Systematic Bacteriology second edition Volume I (Castenholz 2001).
2. Taxonomic scheme according to Cavalier-Smith (Cavalier-Smith 2002).
3. Taxonomic scheme according to the NCBI Taxonomy Browser.

In the Bergey's Manual of Systematic Bacteriology cyanobacteria were classified into five subsections viz. Subsection I (formerly Chroococcales), Subsection II (formerly Pleucapsales), Subsection III (formerly Oscillatorials), Subsection IV (formerly Nostocales) and Subsection V (formerly Stigonemateles). These subsections are distinguished based on the morphological or physiological properties of cyanobacteria such as unicellular/filamentous nature, reproduction by binary fission or multiple fissions.

Cavalier-Smith proposed another taxonomic system in which cyanobacteria are divided into six orders which are Gloeobacterales, Chroococcales, Pleurocapsales, Oscillatoriales, Nostocales, and Stigonematales.

National centre of Biological Information (NCBI) had increased the number of Orders as more and more species of cyanobacteria have been discovered. Currently, all cyanobacteria are divided into nine orders which are Chroococcidiopsidales, Gloeobacterales, Gloeoemargaritales, Nostocales, Chroococcales, Oscillatoriales, Pleurocapsales, Spirulinales and Synechococcales.

Arthrospira platensis is one of the cyanobacteria within the order Oscillatoriales which posses quite a few unique features like tolerance of high pH, halophilic nature and high protein content.

1.2 *Arthrospira platensis* NIES-39

Arthrospira (Spirulina) platensis NIES-39 is an edible cyanobacterium and has been a major attraction due to its multiple uses as feed, dietary supplement, and functional food (Castenholz 2001). It is a rich source of protein (60–70%) (Table 1.1) (Lochab et al. 2014) and other constituents like vitamins, essential amino acids, minerals, and essential fatty acids (Baylan et al. 2012). *Arthrospira platensis* has a higher photosynthetic efficiency hence releasing more oxygen thus producing more food than higher plants. It would serve as nutrients suppliers with their exhaled carbon dioxide and recycled wastes (Oguchi et al. 1987).

Table 1.1 Protein content of various foods. *Spirulina* has the highest protein content.

Food	~Protein content (% dry weight)
Arthrospira platensis (Spirulina)	65
Aphanizomenon flos-aquae	50
Anabaena cylindrica	50
Synechocystis	26
Soybeans	35
Dried milk	35
Peanuts	25
Animal products	15-25
Eggs	12
Grains	8-14
Whole milk	3

Arthrospira (Spirulina) is used as a food source since it was first reported in 1521 A.D. Bernal Diaz del Castillo, a Spanish soldier in the troops of Hernan Cortez's first saw *Spirulina maxima* which was harvested and sold in the local markets of Mexico.

However, people might have used *Spirulina* as food for centuries, though the origin is not known. The next report came from French phycologist P. Dangeard who saw a cake like structure being consumed by the local people of Chad in Africa (Dangeard 1940). These cakes were termed as *dihe* which were later confirmed to be *Arthrospira platensis* by J. Leonard (Leonard 1966, Leonard and Compare 1967).

The first commercial production of *Spirulina* was started in the year 1970 in Lake Texcoco (Gershwin and Belay 2008). Even today, people from Kanembu tribe of Chad consumes *Arthrospira* on a daily basis (Abdulqader et al. 2000). *Spirulina platensis* is now being considered as an edible alga for spacecraft crew in a Controlled Ecological Life Support System (CELSS) (Godia et al. 2002). The World Health Organization thinks that *Spirulina* is an excellent food for human consumption, and *Spirulina* has the Food & Drugs Authority approval for being sold as natural food in the United States.

1.2.1 *Arthrospira platensis* NIES-39: Morphology and Taxonomy

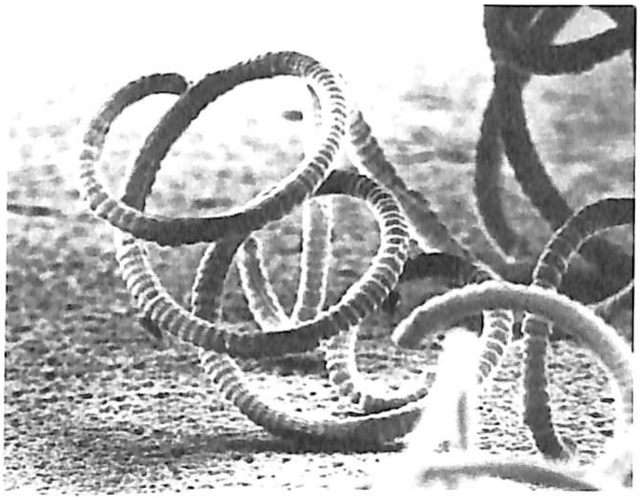


Figure 1.2 Scanning electron micrograph image of a trichome of axenic *Spirulina platensis* (adapted from (Ciferri 1983).

Arthrospira platensis is a filamentous, non-N_2-fixing cyanobacterium which lacks any cellular differentiation like akinete, heterocyst or hormogonium. It is composed of multicellular cylindrical trichomes which are arranged in a left-handed helical filament (Figure 1.2). The filaments are solitary, free-floating and show locomotion through gliding.

The trichomes are covered by a thin sheath, and slight constrictions are present at the cross walls. The width of the trichome ranges from 6 to 12 μm (Geitler 1925). The geometry of the helix could be affected by temperature, chemical and physical conditions (Bai and Seshadri 1983, Bai 1985). However, the helix geometry also depends on different strains of a species. Even variations in trichome geometry have been observed in a natural population (Vonshak 1997). The vegetative cell divides by binary fission in a single plane (Geitler 1925). The cyanobacterial filaments undergo a helical to spiral transition in the solid media (Van Eykelenburg and Fuchs 1980).

This cyanobacterium (*Arthrospira*) is also known as '*Spirulina*' because of its spiral morphology. However, according to the current taxonomic reframes both *Arthrospira* and *Spirulina* belong to two different genera and the genus *Spirulina* is not used as a food supplement (Fujisawa et al. 2010). However, the name *Spirulina* is still used for the trade purposes. The most recent taxonomic evaluation of the species identifies *Arthrospira* as follows (Castenholz 2001) (Table 1.2):

Table 1.2 The taxonomic classification of *Spirulina (Arthrospira) platensis* (Taken from Castenholz 2001).

Kingdom	**Monera** **(Prokaryotae)**	Morphologically simple but metabolically complex and diverse organisms, the bacteria. Lack of a nuclear membrane, and membrane-bound organelles absent – cell division through binary fission – cell simply pinches in two.
Sub-kingdom	**Eubacteria**	'True bacteria' All bacteria that are not archaebacteria are Eubacteria.
Division **(Phylum)**	**Cyanobacteria** **(Cynophyta** **Cyanophytes)** . **Formerly known as** **(Blue-green algae)**	Photoautotrophic bacteria, photosynthesise, but lack chloroplasts. The product of photosynthesis is glycogen and released oxygen. The cells have no flagella or any other type of locomotor organelle. Thylakoids (photosynthetic membranes) are not arranged in stacks. Chlorophyll a, d; blue and red phycobilins, β-carotene, and xanthophylls; storage product, cyanophycean starch; The cell wall is a complex, four-layered structure (consisting of mucopeptides, amino sugars, amino acids and proteins). Gram-negative cell walls ~2500 described species.
Class	**Cyanophyceae**	Single class in cyanobacteria; unicellular or multicellular algae without a true nucleus or chromatophore. Sexual reproduction not known or absent.

Order	**Oscillatorials**	Filamentous, with filament and trichome organisation, hormogones present; heterocysts, akinetes, endospores, hormocysts present; true branching absent, false branching present.
Family	**Oscillatoriaceae**	Filamentous (unbranched); producing hormogonia, many showing a spiral movement by rotation along the longitudinal axis; binary fission; no specialised cells, heterocysts and spores absent; ~1000 species.
Genus	***Arthrospira*** ***(Spirulina)***	Trichomes (filaments) multicellular, cylindrical, without sheath, loosely and regularly coiled (spiralled), usually comparatively short and fewer coils; cross-walls distinct, apices slightly or not all tapering, terminal cell rounded, calyptra absent.
Species	***platensis***	Thallus blue-green; trichomes slightly constructed at the cross-walls, 6-8 mm broad, not attenuated at the ends or only a little attenuated, more or less regularly spirally coiled; spirals 26-36 mm broad, distances between the spirals 43-57 mm; cells nearly as long as broad, or shorter than broad, 2-6 mm long, cross-walls granulated; end-cells broadly rounded.

1.2.2 Genomic structure of *Arthrospira (Spirulina) platensis* NIES-39

The complete genome of *Arthrospira platensis* NIES-39 was sequenced by (Fujisawa et al. 2010). This genome is composed of a single, circular chromosome of 6.8 Mb, without any plasmids. The genome comprises of 6630 protein-coding genes, with 49 RNA genes, two sets of rRNA genes, 40 tRNA genes representing 20 tRNA species. Out of the total 6630 potential protein-coding genes, 5157 (78%) were found orthologous or had similarity to genes of previously known function or other hypothetical genes. However, the remaining 1473 (22%) showed no significant similarity to any known genes.

1.2.3 *Arthrospira platensis*: The current research

Being a successful commercial species, a lot of research has been going on *Arthrospira platensis*. The first area of research on *Arthrospira platensis* focuses on one fundamental question: How to increase the biomass of the commercially produced *Spirulina*. A handful of literature is available on the standardization of various growth conditions/parameters (light, pH, temperature etc.) which could affect the final biomass of this species (Pandey et al. 2010, Ajayan 2011, Godoy Danesi et al. 2011, Mohite and Wakte 2011, Markou et al. 2012, Markou 2015, Devanathan et al. 2016).

Arthrospira platensis is a halophilic cyanobacterium and can sustain in high salt concentrations (>30 g/l) (Vonshak et al. 1988, Zeng and Vonshak 1998, Habib et al. 2008). It is also an alkalophile, i.e. it grows in water with high pH. The optimum pH for the growth of *Arthrospira platensis* is from 8.5-11 (Habib et al. 2008). Therefore, the second area of research is to find out the possible genes/proteins or mechanisms in *Arthrospira platensis* which are responsible for stress tolerance. This cyanobacterium can tolerate such high salinity due physiological mechanisms such as an accumulation of several inorganic and organic osmoregulators (Reed et al. 1986, Warr et al. 1988) and also by the active expulsion of sodium ions from the cell (Gabbayazaria et al. 1992, Peschek et al. 1994). It has been shown that carbohydrates metabolism in *Arthrospira platensis* cells increases during adaptation to salinity (Warr et al. 1985, Vonshak et al. 1988). All the cyanobacteria contain some substances like sucrose, trehalose, glucosyl glycerol or glycine-betaine which will protect membrane and cellular proteins against salt stress (Reed et al. 1986, Ferjani et al. 2003). During stress, there is an accumulation of these substances (Rentsch et al. 1996, Kempf and Bremer 1998).

Another possible mechanism of adaptation to high salinity as well as high pH is the exclusion of Na^+ ions from the cells (Apse and Blumwald 2002, Wutipraditkul et al. 2005). Na^+/H^+ antiporters are present in all the cyanobacteria and facilitate the exchange of Na^+ and H^+ across the membrane (Blumwald et al. 2000, Padan et al. 2001, Serrano and Rodriguez-Navarro 2001). In cyanobacterial cells, the active export of Na^+ and accumulation of K^+ is involved in salt adaptation mechanism (Gabbayazaria et al. 1992, Ritchie 1992). Na^+/H^+ antiporters help in Na^+ efflux and prevent the toxic effects of high cytoplasmic Na^+ levels. Na^+/H^+ antiporters are also known to enhance the bacterial growth under alkaline conditions due to acidification of the cytoplasm relative to the external environment (Padan et al. 2005). The role of unsaturated fatty acids in thylakoid membranes is also important as they help the photosynthetic machinery during salt stress (Allakhverdiev et al. 2001).

Researchers (Wang et al. 2013) have also identified 141 differentially expressed proteins of *Arthrospira plantensis* under salt-stress conditions which belong to different pathways like glucose metabolism, photosynthesis, lysine synthesis, cysteine and methionine metabolism, glutathione metabolism, fatty acid metabolism, heat shock protein and ABC transporter. All these pathways help *Arthrospira platensis* against high salt stress.

Another important field of research in *Arthrospira platensis* is to deduce the molecular mechanism behind its high protein content. As already mentioned, that it is used as a food supplement particularly as a protein source, it is crucial to know about the molecular phenomena lying behind the production of high protein content. However little is known about the molecular basis of its high protein production.

1.3 The protein translational machinery

In the gene expression process, the information available in the DNA sequence is used for the synthesis of gene products that could either be a protein or non-protein molecules like RNA molecules (tRNA, rRNA or snRNA). The process of translation consists of many steps involving an array of machinery (Figure 1.3). Each of these steps is regulated by different regulatory mechanisms and hence determine the final protein content of the cell. Any changes in this regulation might lead to severe consequences which can affect the proper functioning of the protein. However, on a different note, this regulation can also help the cell in gaining additional benefits regarding protein production efficiency. Some of these processes that can affect the protein production efficiency are degradation of mRNA, tRNA synthetase and nitrogen acquisition.

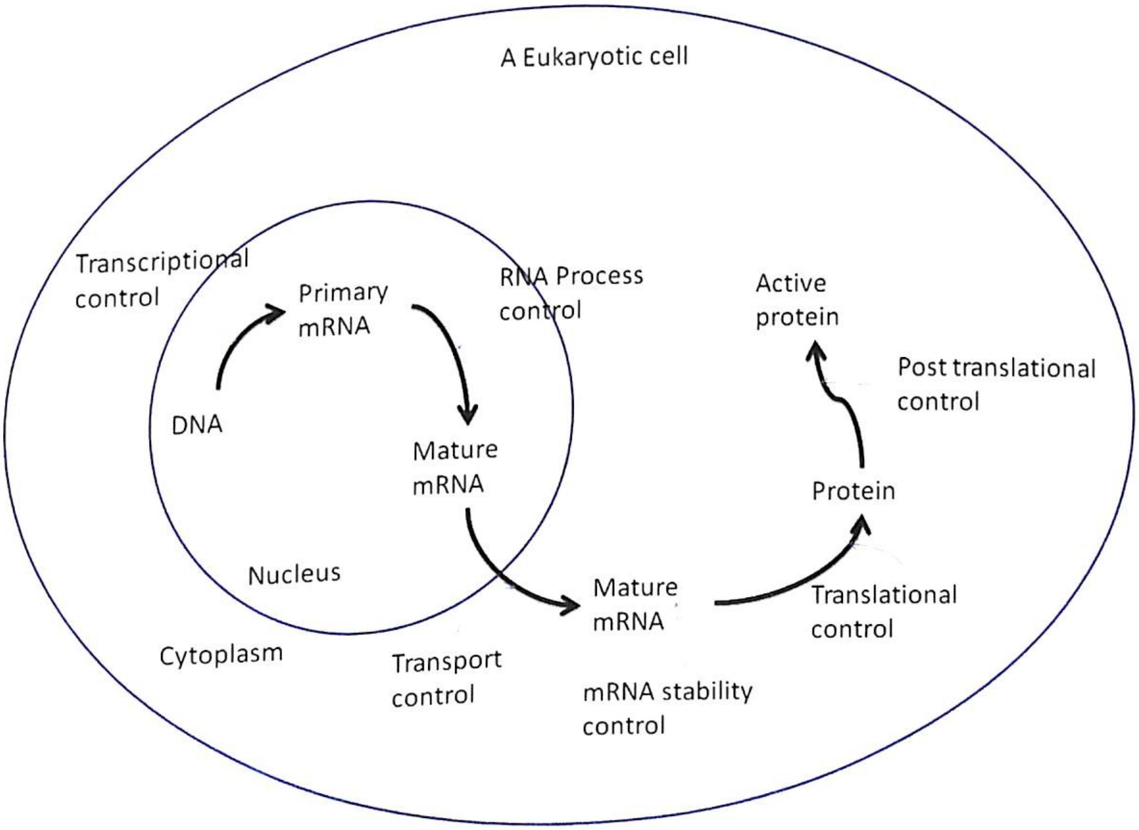


Figure 1.3 The process of making an active protein from a gene. There are different checkpoints in the protein manufacturing process that ultimately affects the final protein content in a cell.

1.3.1 Degradation of mRNA

Degradation of mRNA is a crucial process that enables organisms to rapidly change the pattern/amount of protein synthesis in a changing environment. It directly affects protein synthesis by varying the amount of mRNA available for translation.

The first process that affects the amount of mRNA in a cell is its degradation by Endonucleolytic enzymes. Endonucleolytic cleavage of a polycistronic mRNA can generate different transcripts with different half-lives (Laalami and Putzer 2011). As a result, different amounts of protein can be produced as needed by the organism (Burton et al. 1983, Meinken et al. 2003). However, in another study, mutants for endonucleolytic enzymes were generated to see that the levels of more than 650 transcripts were altered and the relative amount of more than 200 proteins were significantly changed (Mader et al. 2008). It has also been seen that depletion of endonucleolytic enzymes increases the half-life of bulk mRNA more than two-fold (Shahbabian et al. 2009).

The second process which affects the mRNA concentration is its own 5` end. The 5` end is an important site in many prokaryotic mRNA transcripts. The stability of the mRNA transcript depends on the structure and the phosphorylation state of this 5` end. However, different pathways and enzymes are involved in different prokaryotic phyla. By replacing this 5` UTR of a short-lived mRNA with 5` UTR of a highly stable mRNA like *ompA* (15-20 min half-life), it has been observed that the half-life of the recipient has increased compared to that of the donor (Belasco et al. 1986).

1.3.2 Aminoacyl tRNA synthetase

Aminoacyl tRNA synthetase (aaRS) or tRNA ligase is an enzyme that attaches an amino acid to its particular tRNA. The resultant molecule is known as aminoacyl-tRNA. This process is also known as "charging" the tRNA. This charged tRNA carries the amino acid at the translational site. The translation process is highly dependent on the availability of different molecules like mRNA and tRNA. It has been shown that any changes in the concentration of these molecules highly influence the quality and quantity of the resultant protein (Kudla et al. 2009, Rosano and Ceccarelli 2009, Zhang et al. 2009, Gingold and Pilpel 2011, Plotkin and Kudla 2011, Fedyunin et al. 2012).

One of the mechanisms that can influence the final protein product is the availability of tRNA (Dong et al. 1996, Dittmar et al. 2005). The concentration of a particular tRNA/total tRNA can influence the rate of protein synthesis. But again, it has been shown that under stressful conditions tRNA availability can significantly vary between different conditions and also over time (Dong et al. 1996, Dittmar et al. 2005). This variability in the tRNA concentrations can significantly affect the translational process in many ways resulting in a corresponding effect in the protein concentration in the cell (Sorensen et al. 2005, Zouridis and Hatzimanikatis 2008, Wohlgemuth et al. 2013).

1.3.3 Nitrogen acquisition

Protein manufacturing through the translation process relies on the availability of amino acids. These amino acids are synthesised through their anabolic mechanisms (Nelson and Cox 2017). However, these anabolic pathways again are dependent on the availability of nitrogen in the cell as nitrogen is an integral element in the amino acids.

1.3.3.1 Nitrogen

Nitrogen is the most abundant element in the earth's atmosphere with 78.1% present as its dimeric form N_2. Nitrogen is present in many compounds like nitric acid, ammonia, cyanide and organic nitrates. Nitrogen is necessarily present and required by the majority of the living organisms ranging from unicellular prokaryotes to multicellular eukaryotes. It is primarily present in proteins, DNA, RNA and also in ATP. There are a number of ways in which this nitrogen can be acquired depending on the individual and the environment. Different chemical forms of nitrogen are present in the environment which includes nitrate (NO_3^-), organic nitrogen, nitrite (NO_2^-), ammonium (NH_4^+), nitric oxide (NO) nitrous oxide (N_2O), or inorganic nitrogen gas (N_2).

1.3.3.2 Nitrogen cycle

The nitrogen in the atmosphere keeps moving from the atmosphere to biosphere and other organic compounds and again into the atmosphere. This cyclic movement is known as the nitrogen cycle (Galloway et al. 2004, Kuypers et al. 2018) (Figure 1.4). In this cyclic movement, nitrogen is converted from one form to another by both physical and biological processes. Four important processes involved in the nitrogen cycle are (1) Fixation (2) Ammonification (3) Nitrification (4) Denitrification.

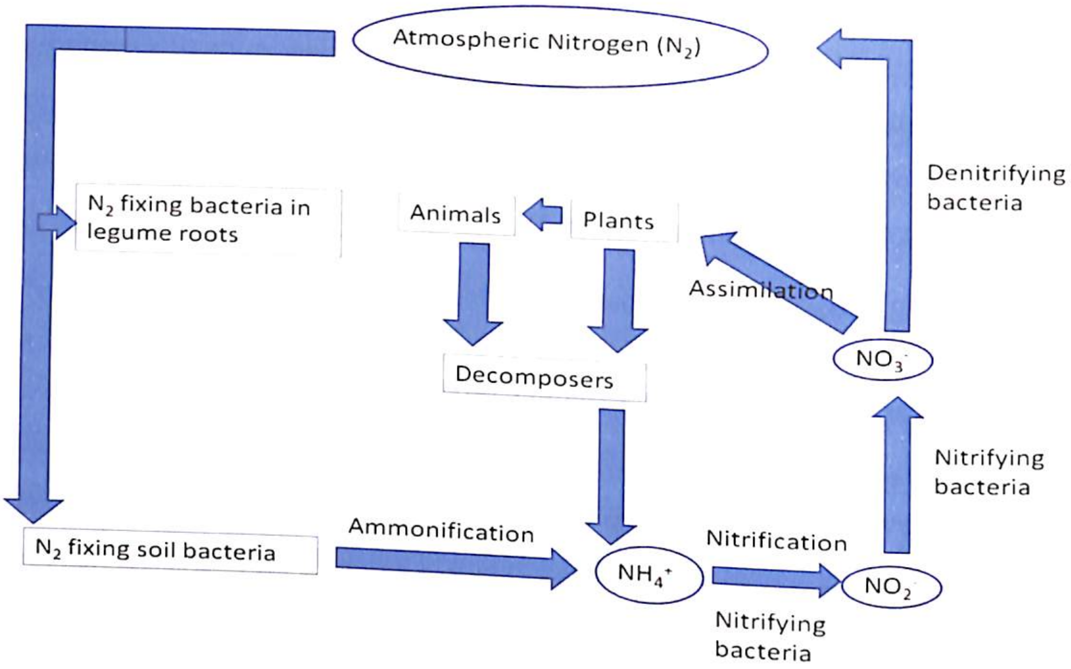


Figure 1.4 Nitrogen cycle: first the nitrogen in the atmosphere is fixed into ammonium by diazotrophic bacteria using the enzymes nitrogenases. Ammonia is absorbed by plants for protein synthesis. Animals eat plants and make their proteins and excrete nitrogen-bearing waste. After the death they undergo decomposition returning free dinitrogen to the atmosphere.

As the most abundant form of nitrogen is the dinitrogen (N_2) present in the atmosphere, it is the most readily available form. However, the triple bond of nitrogen is difficult to break, which makes it difficult to use and only a few specialised enzymes can 'fix' this dinitrogen into the usable form. Nitrogen fixation is a process by which dinitrogen is converted into more usable forms like ammonia (NH_3) or other molecules which could easily be used by other living organisms (Postgate 1983). Naturally, this fixation can be achieved by two methods *viz.* non-biological method and the biological method. In the case of the non-biological method, in the presence of lightning, the atmospheric nitrogen combines with oxygen to form nitrogen oxides (NO_X). This nitrogen oxide can be converted into nitric acid or nitrous acid with the reaction of water and gets into the soil to make nitrate and used by plants.

Biological nitrogen fixation (Hellriegel and Wilfarth 1888) (BNF) is a process where ammonium is produced from atmospheric nitrogen by diazotrophic bacteria using enzymes called nitrogenases (Postgate 1983).

Diazotrophs are a group of bacteria and archaea that fix atmospheric nitrogen gas into a more usable form such as ammonia (Puri et al. 2015, Padda et al. 2016, Puri et al. 2016, Puri et al. 2016). These diazotrophs depending on their habitat can be divided into two categories viz.

free-living and symbiotic. Free-living diazotrophs are those that are living freely in a medium like soil, water etc. Examples of these free-living bacteria include *Clostridium*, *Desulfovibrio*, *Methanococcus*, *Klebsiella pneumoniae*, *Paenibacillus polymyxa*, *Azotobacter vinelandii* etc. Cyanobacteria are present in almost all the environments of the earth and play a significant role in the carbon and nitrogen cycle. Some cyanobacteria are diazotrophic in nature (Latysheva et al. 2012). Cyanobacteria fix Nitrogen in a coral reef which is about twice as on land. The colony-forming marine cyanobacterium *Trichodesmium* is a highly efficient nitrogen fixer. It fixes about half of the nitrogen in the marine system over the globe (Bergman et al. 2013).

On the other hand, the symbiotic diazotrophic bacteria are associated with some plants species. Examples of these types of bacteria include Rhizobia which are associated with the plants from the legume family. There are also examples of symbiotic cyanobacteria. They are known to have some association with fungi known as lichens, with liverworts, with a fern, and with a cycad (Postgate 1983). These do not form nodules, but they have a specialised cell called heterocyst which excludes the oxygen. The association with fern is important agriculturally: the water fern *Azolla* harbouring *Anabaena* is an important green-manure for rice culture (Postgate 1983).

The next process in the nitrogen cycle is the ammonification where any form of organic nitrogen either from animal waste or dead animal or plant is converted into ammonium. This process involves the decomposers that may be bacteria or fungi.

The third process in the nitrogen cycle is nitrification. In this process, the ammonium generated from the above two processes, i.e. nitrogen fixation and ammonification get sequentially oxidised into nitrite and then to nitrate with the help of the nitrifying bacteria. Ammonium is converted to the nitrite either by bacteria (*Nitrosomonas* and *Nitrosococcus*) or by archaea (*Nitrosopumilus maritimus* and *Nitrososphaera viennensis*). In the second step, nitrite is converted into nitrate mainly by bacteria of the genus *Nitrobacter* and *Nitrospira*.

The nitrate produced in the nitrification process can either be converted back to atmospheric dinitrogen by the process of denitrification. In the denitrification process, bacterial species such as *Pseudomonas* and *Clostridium* anaerobically reduces nitrate to dinitrogen which completes the nitrogen cycle. Or on the other hand, it can be absorbed by the non-nitrogen fixing organisms for their nutritional requirements by the process of assimilation which we will discuss in detail.

The only pathway that can incorporate nitrogen into amino acids is the nitrogen assimilation pathway. Few studies have shown the role of nitrogen assimilation pathway in high protein content of cyanobacteria (Jha et al. 2007, Ali et al. 2008, Lochab et al. 2009). In a comparative study, it has been shown that the *Arthrospira* nitrate-assimilating enzymes (NR, NiR and GS) have higher specific activities and are more stable than those of rice (Jha et al. 2007, Ali et al. 2008, Lochab et al. 2009). Again, in a comparative study between *Arthrospira* and rice, it has been shown that assimilatory enzymes (NR, NiR and GS) of *Arthospira platensis* are more thermotolerant than those of rice (Lochab et al. 2009).

In the assimilation process, the very first step is the intake of nitrogenous compounds inside the cell. This intake is facilitated by various transporters. These transporters include ammonium transporters, nitrate/nitrite transporters or even urea transporters in some cases.

1.4 Nitrogen transport

Different molecular forms of nitrogen like ammonia, nitrite or nitrate are available for nitrogen uptake during the nitrogen cycle. Hence organisms which rely on these molecules (non-nitrogen fixing organisms) have special transporters which help in the intake of these molecules. The major transporters have been described here.

1.4.1 Ammonium transporters

For ammonium transportation, cells have ammonium transporters called Amt proteins (Ammonium transporters). These are structurally related integral membrane proteins. They are found in both plants and bacteria (Khademi et al. 2004, Zheng et al. 2004, Khademi and Stroud 2006). These ammonium transporters are helpful for species (both prokaryotic and eukaryotic) which are found in an anaerobic condition like grasslands (Jackson et al. 1989) or flooded areas like rice fields (Ishii et al. 2011).

1.4.2 Nitrate/nitrite transporters

Nitrate is the major source of nitrogen for many photosynthetic organisms including cyanobacteria, algae and plants (Guerrero et al. 1981). Hence it becomes necessary for these organisms to have at least one active nitrate transporter. Two families of transporters are involved in the nitrate/nitrite transportation. These two families are ATP binding cassette (ABC) type transporters and the Major Facilitator Superfamily (MFS) transporters.

1.4.2.1 ATP-binding cassette transporters

ATP-binding cassette transporters (ABC transporters) is a superfamily of transporters which is present in all the organisms ranging from prokaryotes to higher plants and even humans (Jones and George 2004, Ponte-Sucre 2009). They can be divided into three main categories *viz.* importers who are found only in prokaryotes, exporters who are found in both prokaryotes and eukaryotes and a third category which is involved in DNA repair and translation (Davidson et al. 2008, Goffeau and De Hertogh 2013). The importers transport a wide range of molecules including nutrients, biosynthetic precursors, trace metals and vitamins while the exporters are involved in the transport of lipids, sterols, drugs, and metabolites. One kind of Nitrate/Nitrite transporter belongs to these ABC transporters. These ABC type nitrate transporters (NRT) are found in freshwater species of cyanobacteria and some heterotrophic bacteria (Omata et al. 1993, Wu and Stewart 1998). The cyanobacterial ATP-binding cassette (ABC) type permeases are involved in nitrate uptake (Flores and Herrero 2005). It consists of a periplasmic membrane-adhered substrate-binding protein, and in the cytoplasmic side, it contains two transmembrane subunits and two ATPase subunits. Two cytoplasmic subunits of ABC-type uptake transporter power the transport reaction and are highly conserved throughout cyanobacterial genera (Flores et al. 2005). They were initially identified in *Synechococcus elongatus* (Madueño et al. 1988, Omata et al. 1989, Sivak et al. 1989, Sazuka 2003).

In cyanobacteria, ABC type NRT are encoded by *nrtABCD* genes (Omata et al. 1993) (Figure 1.5). It is a bispecific transporter which transports both nitrite and nitrate with a high affinity (Luque et al. 1994, Maeda and Omata 1997). This transporter contains four polypeptide chains which are NrtA, NrtB, NrtC and NrtD. NrtA which is a high affinity periplasmic solute-binding lipoprotein searches for nitrate/nitrite as it can bind both nitrate and nitrite when nitrate is the primary nitrogen source (Maeda and Omata 1997). Now NrtA transfers this nitrate/nitrite to NrtB, which is an integral membrane permease. Nitrate/nitrite comes inside the cell through this membrane protein. Cytoplasmic NrtC and NrtD helps in the movement of the nitrate/nitrite molecules across the membrane through ATP hydrolysis as both of them are ATPase and contain ATPase domain. In addition to the ATPase domain, NrtC also has a solute-binding domain and hence it is a fusion protein. NrtC also regulates this transport process (Omata 1995, Kobayashi et al. 1997, Koropatkin et al. 2006). The NrtC shares 50% homology with NrtA.

Spirulina platensis genes for ABC transporters are arranged in an operon (*NrtA-B-C-D*) (Omata et al. 1993, Fujisawa et al. 2010). ATP hydrolysis provides them energy for solute transport across cell membranes. Membrane-spanning domains of the permease undergo conformational changes induced by ATP binding and hydrolysis (Davidson and Chen 2004). NrtA can bind both nitrate and nitrite in the periplasm. NrtA is 440 amino acids long protein and is anchored to the cytoplasmic membrane (Maeda and Omata 1997). NrtB bears six transmembrane segments, which are highly hydrophobic and are about 280 amino acids long (Wu and Stewart 1998). NrtD is also a conserved protein of about 275 amino acids.

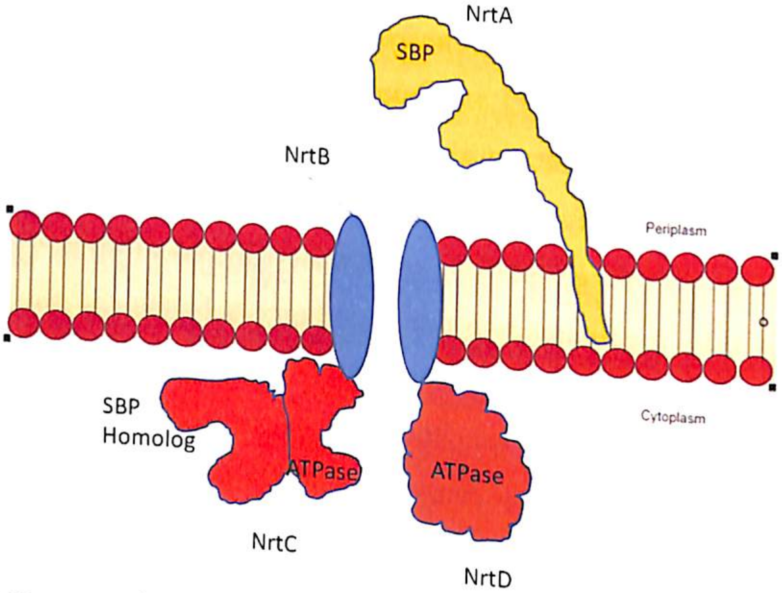


Figure 1.5 Cartoon representation of the NrtABCD nitrate transporter. NrtA is present in periplasmic space and can bind both nitrate and nitrite which are then transferred to NrtB complex. The transmembrane pore is usually formed by a dimer of two transmembrane spanning polypeptides. NrtC and NrtD are ATPases that couple ATP hydrolysis to nitrate and nitrite transport through the pore. NrtC contains a C-terminal solute-binding domain (adapted from Koropatkin 2006).

1.4.2.2 Major Facilitator Superfamily

The major facilitator superfamily (MFS) of membrane protein transport small solutes across cell membranes. They work according to the chemiosmotic gradient (Pao et al. 1998, Walmsley et al. 1998). MFS of transporters is divided into many families. Two major families include peptide transporter (PTR) family of the MFS which contains NRT1 transporter (Figure 1.6) and is found in all vascular plants and the second one is nitrate-nitrite-porter (NNP) family of MFS (Forde 2000, Galvan and Fernandez 2001) which contains the NRT2 transporter and is found in all the eukaryotes, cyanobacteria and heterotrophic bacteria. NRT2 is mainly found in the marine species of cyanobacteria. NRT2 like transporter is encoded by the *nrtP* genes (Sakamoto et al. 1999), and it is also found to be bispecific (Wang et al. 2000, Allen et al. 2001). The MFS contains different proteins that are about 500-600 amino acids in length. They have a membrane topology containing two sets of six transmembrane helices which are connected by a cytosolic loop (Henderson 1991, Baldwin 1993, Pao et al. 1998).

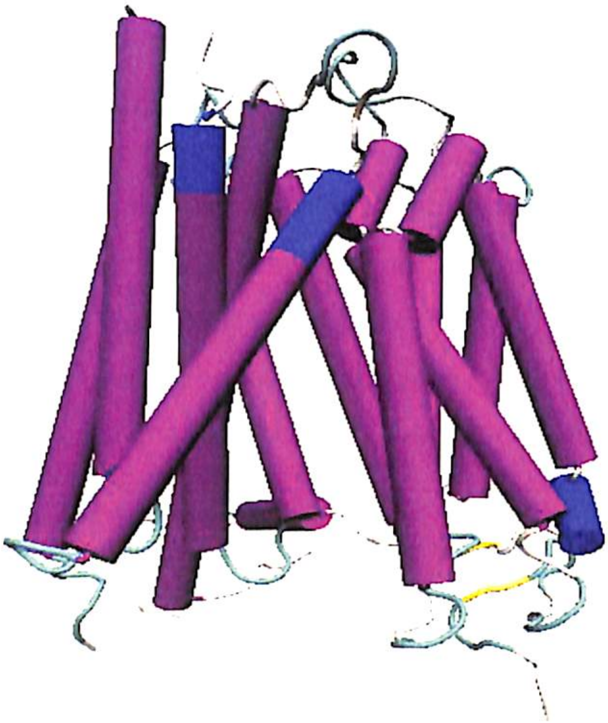


Figure 1.6 Crystal structure of NRT1.1 (PDB – 4OH3). This structure shows 12 transmembrane helices. Two chains are present in this structure, and only one chain is shown here.

1.5 Nitrogen assimilation

Nitrogen assimilation is one of the major processes of nitrogen acquisition in cyanobacteria (Figure 1.7). The nitrogen assimilation process in cyanobacteria is initially described by (Guerrero et al. 1981). Most of the cyanobacteria absorb nitrate through transporters and assimilate this nitrate via assimilation pathway (Herrero et al. 2001, Garcia-Fernandez et al. 2004, Ohashi et al. 2011). In the process, nitrate is transported into cells by an active transport system, and this absorbed nitrate (NO_3^-) sequentially gets reduced to NH_4^+ by two enzymes *viz.* assimilatory nitrate reductase (NR-1.7.1.1-3) and nitrite reductase (NiR-1.7.1.1) respectively. Generated ammonium then enters the GS-GOGAT pathway (Merrick and Edwards 1995, Reitzer 2003). In this pathway, there are two enzymes *viz.* Glutamine

Synthetase (GS-6.3.1.2) and Glutamate Synthase (GOGAT-1.4.7.1) which helps in incorporating the nitrogen initially into Glutamate and Glutamine and hence to the rest of the nitrogen-containing molecules.

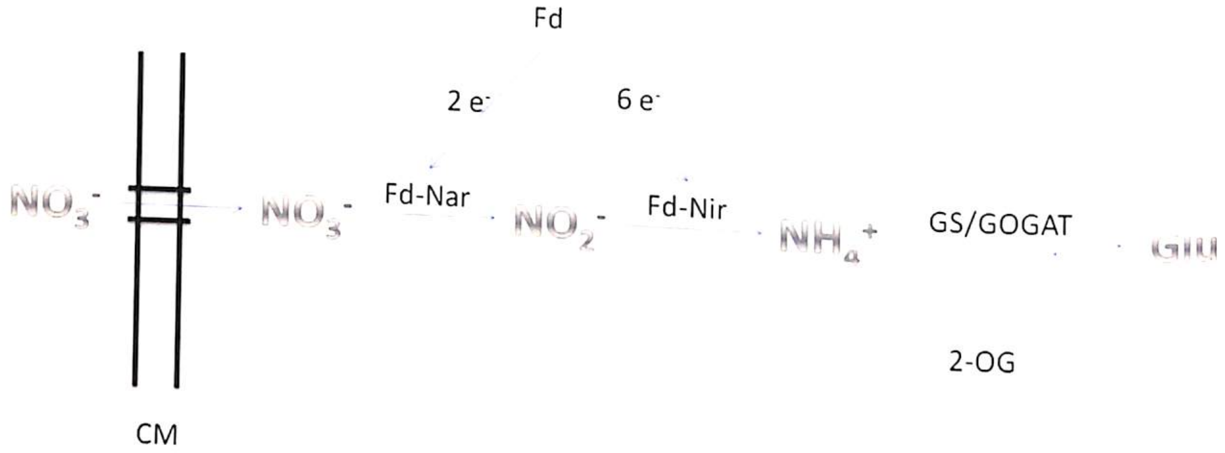


Figure 1.7 The nitrate assimilation system of fresh-water cyanobacteria like *Synechococcus elongatus* or *Anabaena* sp. Strain PCC 7120. The process starts with the intake of nitrate by ABC-transporters and then the sequential reduction of nitrate to ammonium through nitrate assimilation enzymes like nitrate reductase, and nitrite reductase. The resulting ammonium enters the glutamine synthetase/glutamate synthase (GS/GOGAT) pathway and finally in the amino acid anabolism.

1.5.1 Nitrate assimilation

1.5.1.1 Nitrate reductase

NR is found in all forms of life ranging from plants, algae, fungi, archaea, and bacteria. (Volkl et al. 1993, Zumft 1997, Ramirez-Arcos et al. 1998, Campbell 1999). All prokaryotic NRs (Nas, Nap, Nar; described below) belong to the dimethylsulfoxide (DMSO) reductase family (Hille 1996). Cyanobacterial nitrate reductases are molybdoenzymes that catalyse the reductional conversion of nitrate to nitrite. They can be classified into three groups based on their localisation and function: (i) Respiratory NR (NAR) which are generally present as integral membrane protein complexes and generate the metabolic energy by using nitrate as a terminal electron acceptor. (ii) Periplasmic NR (NAP) are present in the periplasm and help in the dissipation of the excess reducing power for redox balancing. (iii) Assimilatory NR (NAS) which are present in the cytoplasm and utilise nitrate as a nitrogen source for growth (Richardson et al. 2001, Stolz and Basu 2002). Assimilatory nitrate reductase is the first enzyme in the nitrogen assimilation pathway which helps in the incorporation of nitrogen into the biomass (Lin and Stewart 1998, Campbell 1999). Assimilatory NR of cyanobacteria is a 75 to 80 kDa single polypeptide that contains an iron-sulfur cluster (Ida and Mikami 1983, Mikami and Ida 1984). In cyanobacteria, NR contains the bis-molybdopterin guanine dinucleotide (bis-MGD) as a cofactor and a [3Fe-4S] cluster for electron transportation (Rubio et al. 1998, Rubio et al. 1999, Rubio et al. 2002). Electrons are donated by ferredoxin in the cyanobacterial NR (Mikami and Ida 1984, Rubio et al. 1996, Rubio et al. 2002). The nitrite produced by NR is further reduced to either the end product ammonia or the denitrification intermediate nitric oxide (Figure 1.8).

1.5.1.2 Nitrite reductase

The reaction which catalyses the reduction of nitrite to ammonium is mediated by nitrite reductase (NiR). NiR is also found in all the domains of life, and unlike NR, NiR of prokaryotes and Eukaryotes share high sequence homology (Luque et al. 1993). NiR can be divided into dissimilatory and assimilatory categories. The dissimilatory group is again divided into copper containing and multiheme containing (cytochrome cd1 or cytochrome c). The assimilatory group have siroheme [4Fe-4S] as the metal co-factor (Flores et al. 2005). Based on the electron donor they are either NADH dependent (bacteria) or ferredoxin-dependent (cyanobacteria). Cyanobacterial NiR is a monomer of 52-56 kDa molecular weight. Two prosthetic groups, i.e. [4Fe–4S] cluster and a siroheme are present. Ferredoxin or flavodoxin acts as the electron donor (Manzano et al. 1976). NiR converts nitrite to ammonium by 6 electrons reduction mechanism (Knaff and Hirasawa 1991) (Figure 1.8).

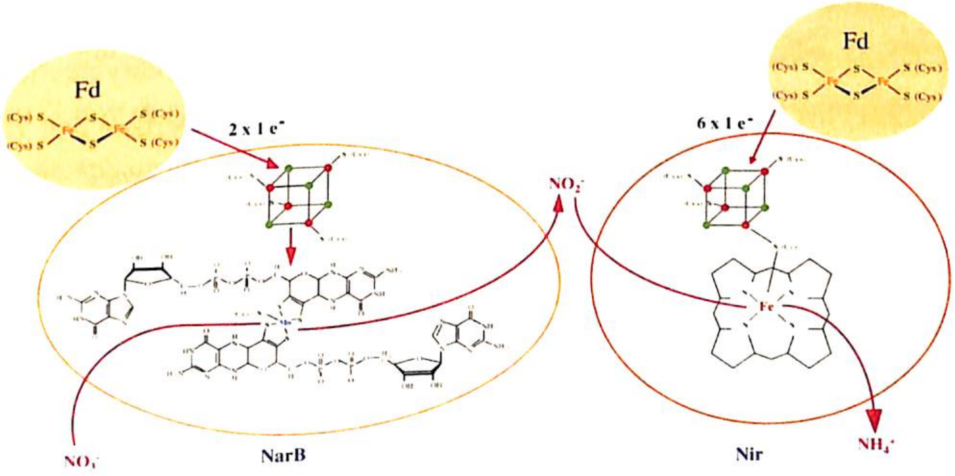


Figure 1.8 Nitrate reductase (NarB) and nitrite reductase (Nir) proteins from *Synechococcus elongatus*, along with their prosthetic groups (iron-sulfur centre and molybdenum cofactor for NarB; iron–sulphur centre and siroheme for Nir) and their interactions with the substrates and ferredoxin (Fd) Iron atoms are in red, and sulfur atoms in green (Adapted from Flores 2005).

1.5.2 GS/GOGAT pathway for ammonium assimilation

GS/GOGAT pathway is the most prevalent pathway in organisms for ammonium assimilation (Merrick and Edwards 1995, Reitzer 2003). There are two enzymes in this pathway which are Glutamine Synthetase (GS) and Glutamate 2-oxoglutarate aminotransferase (GOGAT) also known as Glutamate synthase (Figure 1.9). In cyanobacteria, this pathway has been shown to be the major ammonia-assimilating route (Dharmawardene et al. 1973, Stewart and Rowell 1975, Wolk et al. 1976, Meeks et al. 1977, Rowell et al. 1977).

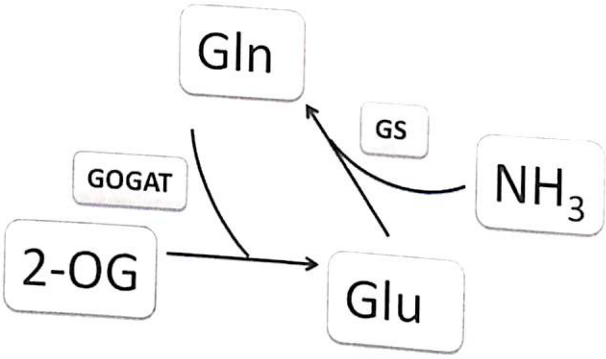


Figure1.9 GS/GOGAT Cycle involving ammonium incorporation using 2-OG carbon skeleton.

1.5.2.1 Glutamine Synthetase (GS)

Glutamine synthetases (GS) (6.3.1.2) are an enzyme family of large oligomeric proteins that catalyse the condensation of ammonia and glutamate to form glutamine. Glutamine is the main nitrogen source for protein and nucleic acid synthesis (Van Rooyen et al. 2011, Saelices et al. 2015). GS is present in all the forms of life ranging from prokaryotes to eukaryotes (Pesole et al. 1991) because it is critical to nitrogen metabolism (Robertson and Tartar 2006). GS has been categorised into three different classes (Kumada et al. 1993, Eisenberg et al. 2000).

1. Class I GS enzymes (GSI) are only found in prokaryotic organisms. They are dodecamers arranged in two rings of 6 each. This is 450 to 470 amino acid long enzyme (Yamashita et al. 1989, Brown et al. 1994).
2. Class II enzymes (GSII) are found in both bacteria (family Rhizobiaceae, Frankiaceae, and Streptomycetaceae) and eukaryotes. GSII is also a multimer of ten identical subunits with 350 to 420 residues (Kumada et al. 1993, Krajewski et al.

2008). In case of plants isozymes of GSII are present in both chloroplast and cytoplasm.

3. Class III enzymes (GSIII) are newly discovered and have only been detected in *Bacteroides fragilis* and *Butyrivibrio fibrisolvens*. It is a dodecamer formed by double-rings of identical chains (Van Rooyen et al. 2011). Their size is about 700 amino acids.

Oligomers of all the classes are arranged into two rings lying face-to-face with each other (Eisenberg et al. 2000, Krajewski et al. 2008).

Talking about prokaryotic GS, they are dodecamers which are arranged in two rings. The two rings of this GS are being held together using hydrogen bonding and hydrophobic interactions (Eisenberg et al. 2000). Each ring contains six monomers. An active site is present between two monomers, and hence a total of 12 active sites are present. Each active site is a funnel like structure in which three distinct substrates namly a nucleotide, ammonium ion, and amino acid would bind (Liaw et al. 1995, Eisenberg et al. 2000, Krajewski et al. 2008). ATP occupies the top position of this bifunnel (Liaw et al. 1993, Liaw et al. 1994, Liaw et al. 1995). Glutamate occupies the bottom position of the active site (middle part of bifunnel) (Liaw and Eisenberg 1994). Space between the nucleotide and the amino acid binding site is the place where divalent cations (Mn^{+2} or Mg^{+2}) bind. These cations help in the transfer of the phosphoryl group from ATP to glutamate, and it also provides stability to GS and helps in the binding to glutamate (Eisenberg et al. 2000).

Cyanobacteria contain class I GS which is a homo dodecamer with 12 active sites where the molecular weight of each subunit is ≈ 55 KDa (Eisenberg et al. 2000).

GS combines glutamate with ammonia to yield glutamine through an ATP-dependent condensation (Liaw et al. 1995). The hydrolysis of ATP is the first step in this process. ATP transfers its phosphate to glutamate to form an intermediate which is γ-glutamyl phosphate. This intermediate reacts with ammonium to form the final product glutamine and inorganic phosphate. Only after the glutamine is released ADP and P_i dissociate. Glutamine dissociates from the enzyme's active site through its bottom while the inorganic phosphate leaves the active site from the top (Hunt et al. 1975).

1.5.2.2 Glutamine 2-OxoGlutarateAminoTransferase (GOGAT)

Glutamate synthase (glutamine: 2-oxoglutarate aminotransferase [GOGAT]) is the most important enzyme in the nitrogen assimilation pathway. This enzyme transfers the amide group of glutamine to 2-oxoglutarate and hence producing two molecules of glutamate (Forde and Lea 2007). GOGATs are classified into two classes based on their electron donors (Vanoni and Curti 1999). The first class of GOGAT derives its electron from NADPH. This NADPH GOGAT is unique to bacteria and is often called as "bacterial GOGAT". The second type of GOGAT is ferredoxin-dependent (Fd-GOGAT) and uses the ferredoxin coming from photosynthesis as an electron donor. This type of Fd-GOGAT is found only in chloroplasts of plants and cyanobacteria, and hence it is also known as "plant type GOGAT".

Cyanobacterial GOGAT is a monomeric protein of 50 kDa while bacterial-GOGAT is a hetero-octamer. Fd-GOGAT and the alpha subunit of NADPH-GOGAT are homologous to each other (Kameya et al. 2007). Four domains are present in both Fd-GOGAT and the alpha subunit of NADPH-GOGAT. The first one is the glutamine amidotransferase (GATase) domain at which glutamine is hydrolysed, and ammonium is generated. The second is the central domain which connects the GATase domain and the synthase domain. The ammonium generated at the GATase domain gets translocated to the third domain which is the synthase domain via an intramolecular ammonia channel. This channel helps the enzyme in binding the ammonium, and both central and fourth α-helical domain's residues contribute to this channel. At the synthase domain, ammonium reacts with 2-OG to produce two molecules of glutamate (Kameya et al. 2007).

1.5.3 Regulation of Nitrogen assimilation

1.5.3.1 Regulation through NtcA protein

NtcA is a dimer of two monomeric units (~222 amino-acid, Figure 1.10) present in almost all cyanobacteria (Herrero et al. 2001, Zhao et al. 2010). The determined crystal structure of NtcA has given an insight into its mode of action (Figure 1.10) (Zhao et al. 2010). The transcriptional activity of NtcA is regulated by the binding of an effector molecule (2-OG) to the N-terminal effector binding domain (EBD) (Figure 1.10). The 2-OG binds to the EBD at a pocket which is similar to that used by cAMP in catabolite activator protein, but with a different pattern (Zhao et al. 2010). When 2-OG binds to the EBD, the binding affinity of NtcA towards NtcA promoter increases (Kolb et al. 1993). Structural analysis has revealed that a tighter coiled-coil conformation of the two C-helices of NtcA, induced by 2-OG maintains the proper distance between the two F-helices for DNA recognition (Zhao et al. 2010). NtcA activates the expression of all the genes of nitrogen assimilation including the *nir* operon (Vega-Palas et al. 1990, Vega-Palas et al. 1992, Frias et al. 1994, Luque et al. 1994, Luque et al. 2004). NtcA mediated regulation of nitrogen control depends on modifications of both enzyme activity and gene expression (Herrero et al. 2001).

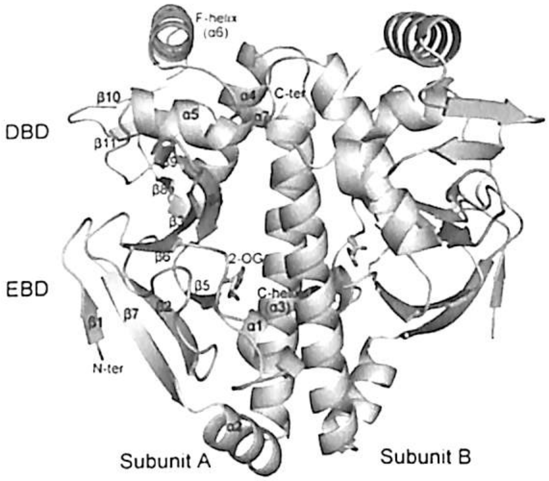


Figure 1.10 Overall structures of NtcA homodimer with 2-OG. The secondary structure elements are numbered sequentially (Adapted from Zhao et al. 2010).

1.5.3.2 Regulation through P_{II} protein

The next level of control in nitrogen assimilation in cyanobacteria is mediated by a signal transduction protein, P_{II} (Burillo et al. 2004). They are another central molecule for perception and signalling of the cellular nitrogen status, recognising ATP and 2-OG (Little et al. 2000, Smith et al. 2003, Burillo et al. 2004, Forchhammer 2004). ATP and 2-OG control the reactivity of P_{II} towards various targets (Jiang and Ninfa 1999, Little et al. 2002). Phosphorylation at Ser49 in response to the cellular nitrogen and carbon supply is the key factor determining its activity (Figure 1.11). Elevation in 2-OG levels signals this dephosphorylation (Forchhammer and Tandeau de Marsac 1995, Irmler et al. 1997). Dephosphorylation of P_{II}-P depends on a protein phosphatase, PphA, which is highly sensitive to 2-OG (even in sub-millimolar range) (Ruppert et al. 2002, Forchhammer 2004). Presence of ammonium initiates dephosphorylation. Presence of a nitrogen source induces medial P_{II} phosphorylation, which is modulated by the inorganic carbon supply to the cells. Nitrogen starvation induces the highest levels of P_{II} phosphorylation (Figure 1.11) (Forchhammer and Tandeau de Marsac 1995, Forchhammer 2004).

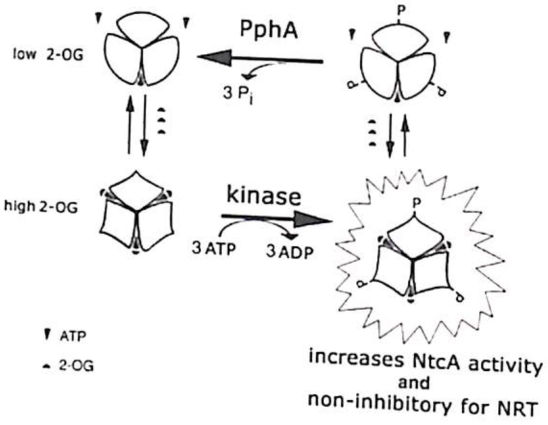


Figure 1.11 P_{II} phosphorylation cycle in response to cellular 2-oxoglutarate levels (Adapted from Forchhammer 2004).

P_{II} signalling mediates the NtcA activated gene expression under conditions of nitrogen starvation (Fadi Aldehni et al. 2003, Paz-Yepes et al. 2003). However, other direct targets of interaction with P_{II} are still to be revealed. An N-acetyl-l-glutamate kinase (NAGK) was recently identified as one of the targets of P_{II} signaling (Burillo et al. 2004, Heinrich et al. 2004). NAGK, the key enzyme in arginine biosynthesis, forms a tight complex with non-phosphorylated P_{II}, enhancing the catalytic activity of this enzyme (Heinrich et al. 2004, Maheswaran et al. 2004).

1.6 Gaps in the existing research

Being an economically important cyanobacterium, *Arthrospira platensis* NIES-39 possesses many characteristic features like nitrogen use efficiency, biotic/abiotic stress tolerance and high protein content which makes this microorganism an ideal model system. As it is already discussed that 22% of potential protein-coding genes are un-annotated. The functional annotation of these proteins could enrich our understanding regarding the molecular basis of the observed characteristic features of this organism.

Arthrospira platensis is known for its high protein content (~65% of dry weight). It has been used as a food supplement from ancient times. Nowadays commercial production has also started and hence it gained economic importance. Due to this, a lot of research has focused on improving the biomass in the production process. There are only a few studies which suggest the role of nitrogen assimilation pathway in high protein content of *Arthrospira platensis* (Jha et al. 2007, Ali et al. 2008, Lochab et al. 2009). The big question is how *Arthrospira platensis* steers its nitrogen assimilation pathway for high protein production? Although it has been shown that *Arthrospira's* nitrate-assimilating enzymes (NR, NiR and GS) have higher specific activities and are more stable than those of rice (Jha et al. 2007, Ali et al. 2008, Lochab et al. 2009). The enzymes (NR, NiR and GS) are also shown to be more thermotolerant than those of rice (Lochab et al. 2009). However, there is an apparent lack of the identification of the molecular basis for the production of high protein content. Thus, these findings further motivated us to look into the sequence and structural features of these enzymes involved in nitrate assimilation of *Arthrospira platensis*. For this study, we have compared all the completely sequenced genomes of cyanobacteria since there is a lack of studies to compare the nitrogen assimilation pathway proteins across the cyanobacterial class. In this study, we have compared the protein sequences and structures between all the completely sequenced cyanobacteria. Sequence plays a key role in determining the function of the protein, and hence any variation in the sequence could affect the function. We have also modelled the protein structures to look into the various changes in the protein domains and fold, various insertion/deletion/substitution in the protein core which can provide us with important clues into its function.

Because *Arthrospira platensis* can serve as "complete food" in itself, understanding how *Arthrospira* has acquired these special abilities will become important and essential to know. These findings will enhance our knowledge of the unique features of the enzymes of the nitrogen assimilation pathway, which may then be extrapolated to agriculturally important crop plants. Thus, the objective of the present works was:

1.7 Objectives of the study

1. Functional annotation of the remaining 22% of the potential protein-coding genes of *Arthrospira platensis* NIES-39 genome.
2. To look into the sequence and structural features of enzymes involved in the nitrogen assimilation of *Arthrospira platensis* NIES-39 for their putative role in the high protein content in the cell.

Chapter II

Methodology

2.1 Overview

This chapter describes the methodology that we have used for our studies with two broad objectives: annotation of the unannotated proteins of *Arthrospira platensis* NIES-39 genome and elucidation of the possible role of nitrogen assimilation proteins in the high protein content of *Arthrospira platensis* NIES-39. The genome of *Arthrospira platensis* NIES-39 had been sequenced and annotated in 2010 (Fujisawa et al. 2010). A total of 6630 protein-coding genes along with 49 RNA genes and 40 tRNA genes were identified. However, Fujisawa et al. could only be able to annotate 5157 (78%) of the genes while rest of 1473 (22%) are still un-annotated. This is to be noted that the 78% of annotated genes also included the protein sequences which were homologous to another hypothetical protein. So, the number of protein sequences with no functional verification is more than 1473. In the study, we have functionally annotated the hypothetical protein sequences of *Arthrospira platensis* NIES-39 using the available online tools and databases. Figure 2.1 depicts the general methodology used for the annotation process.

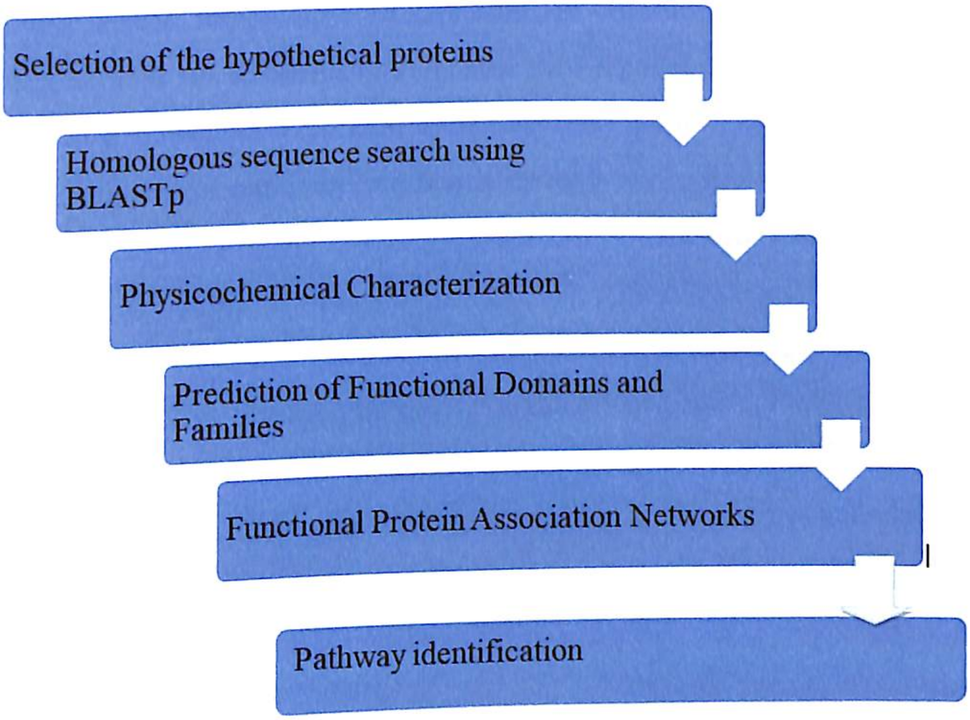

Selection of the hypothetical proteins
Homologous sequence search using BLASTp
Physicochemical Characterization
Prediction of Functional Domains and Families
Functional Protein Association Networks
Pathway identification

Figure 2.1 Overview of the annotation procedure adopted in this study.

For our second objective, fully sequenced cyanobacterial genomes within the NCBI database were selected. Nitrogen assimilation pathway protein homologs of all the selected cyanobacterial species were retrieved using *Arthrospira platensis* NIES-39 protein sequence as a query. Different database searching tools were used for homology searching.

Functionally important residues were identified using sequence comparison. Functionally conserved domains and motifs were identified using CD search and MEME suite. Conservation patterns of conserved residues were identified using Weblogo. Both neighbor-joining (NJ) and maximum likelihood (ML) methods were used to construct the phylogenetic trees. If the rate of evolution between different taxa is not constant NJ is a better choice. ML is more accurate when the species under study are more diverse in terms of their evolution. Speciation and duplication event among the species were inferred using the integrated algorithm in MEGA7.

Species tree based on 16sribosomal RNA gene sequences, gene tree based on the homologous gene sequences of nitrogen assimilation pathway proteins and also the protein tree based on respective homologous protein sequences were generated using both NJ and ML methods. All the above phylogenetic trees were compared among each other to understand the evolution of nitrogen assimilation pathway proteins and to detect any possible mechanism that can explain the high protein content of *Arthrospira platensis* NIES-39.

Structural studies were performed on selected species from major orders. Structures were generated using homology modeling which were further analyzed for their structural quality. Possible structural changes in these modelled structures were identified which could tell us about the evolutionary pattern of these proteins and provide clues to the high protein content of *Arthrospira platensis* NIES-39.

2.2 Selection of the hypothetical proteins for annotation

A complete list of hypothetical proteins of *Arthrospira platensis* NIES-39 was downloaded from the NCBI genome database (genome ID = 171004). However, due to redundancy, these hypothetical proteins were again screened using the manually curated UNIPROT database, and only the non-redundant sequences were used for the annotation process.

2.3 Sequence retrieval and analysis

To find close homologous, these sequences were searched against protein databases using the default parameters of BLASTp (Matrix = BLOSUM62, Word size = 10, Gap Cost = Existence:11 Extension:1) (Altschul et al. 1990). BLASTp is a fast and reliable online tool to find sequence similarity. Only significant hits with identity > 40%, query coverage > 50% and e-value < 0.005 were considered as close homologous. BLASTp is an online program which searches for the similar sequences for a protein query sequence within other online databases. It performs a local alignment between the query and the sequences in databases and shows those alignments. There are different versions of BLAST, like Protein BLAST that compares a protein query to protein databases, Nucleotide BLAST that compares a nucleotide query to nucleotide databases, blastx that translate a nucleotide query in all six reading frames and then compares it to protein databases and a tblastn that compares a proteins query to translated nucleotide databases. The reliability of the hits from BLAST is assessed in terms of an E-value. This is the expected value which tells how many results we would get by chance at any particular score.

2.4 Physicochemical Characterization

ExPASy–ProtPram server (Gasteiger et al. 2003) was used to calculate different physiochemical properties such as isoelectric point, molecular weight and grand average of hydropathicity (GRAVY) of all the hypothetical proteins. The GRAVY value is calculated by adding the hydropathy value for each residue and dividing by the length of the sequence (Kyte and Doolittle 1982). Increasing positive score indicates a greater hydrophobicity. It can tell us about the possible working environment of the protein. For example, high values indicate that the protein has more hydrophobic residues and could be a membrane spanning protein.

2.5 Prediction of Functional Domains and Families

Conserved Domain (CD)-search tool of NCBI (Marchler-Bauer and Bryant 2004) was used for domain identification. This tool searches a comprehensive collection of domain models using BLAST heuristics and imports the domains from different available domain databases like conserved domain database (CDD) (Marchler-Bauer et al. 2015), NCBI curated database, SMART (Letunic et al. 2004), PFAM (Bateman et al. 2004), Clusters of Orthologous Groups (COGs) (Tatusov et al. 2003) and TIGR (Haft et al. 2001). CD-Search tool provides a comprehensive result in terms of specific hits, non-specific hits and superfamilies.

2.6 Functional Protein Association Networks

Proteins mainly work in networks. The interactions between the proteins mainly define their activity and function. Different proteins/enzymes coordinate with each other to regulate a process/mechanism. Understanding these interactions will give us useful insights into protein functioning. The STRING database (version 10.0) was used to predict protein interacting partners for the hypothetical proteins. STRINGS database can predict the interactions that are direct (physical) and indirect (functional) associations, experimental or co-expression (Szklarczyk et al. 2017). The confidence of the interaction was set to medium (0.40) with >10 interaction networks.

2.7 Pathway identification

KEGG database was used to identify the putative pathway in which the annotated hypothetical protein could be involved in (Kanehisa and Goto 2000).

2.8 Selection of cyanobacterial species

NCBI genome database (https://www.ncbi.nlm.nih.gov/genome) lists all the cyanobacteria which have been sequenced at a different level of sequencing (complete, chromosome, scaffold and contig). During complete genome-level sequencing, all the chromosomes are sequenced without any gap with an ambiguity of less than ten nucleotides. In this case, all the possible chromosomes of the species are present. Even plasmids are sequenced without any gap. The second is the chromosome level in which the sequence from single or multiple chromosomes is present. This chromosome may or may not have gaps in it. In scaffold level, several contigs have been joined to form the scaffolds which are unlocalized. In contig level sequencing, only sequences of contigs are reported. Out of these four levels, the genomes which were present in the complete and the chromosome levels were considered for our study.

2.9 Retrieval of nitrate assimilation pathway proteins homologs

Nitrate assimilation pathway proteins from *Arthrospira platensis* NIES-39 were used as a query to retrieve the homologs of these proteins from the (NCBI) RefSeq database. RefSeq database contains non-redundant and curated genomic DNA, mRNA, and protein sequences generated by NCBI. RefSeqs provide a good reference point for genome annotation, identification and characterisation of genes, polymorphism and mutational studies, gene expression studies, and comparative analyses. Blastn and Blastp (Basic Local Alignment Search Tool) (Altschul et al. 1990) were used against organism cyanobacteria (taxid 1117) for retrieving the homologous sequences of genes and the proteins from NCBI (E-value $\leq 1 \text{ X } 10^{-5}$). We have also performed PSI-BLAST (Position Specific Iterative BLAST) (Altschul et al. 1997). for identifying distant homologues. This blast carries out multiple iterations of the results obtained by the first round of blast. From the highest scoring results of the first round, it makes a multiple alignment and then calculates a matrix which is the Position Specific Scoring Matrix (PSSM). This PSSM stores the conservation patterns of the homologous sequences as the score. Now in the second round of PSI-BLAST, this PSSM is used as an input to find more homologues. After the second round, identified new homologous sequences (above the threshold) were added to the matrix and this process iterates for the specified number of times or until no new significant sequences are added to the matrix. This method is more useful in identifying distant homologues.

In addition to the gene and protein sequences of the proteins of study, we also downloaded the 16s rRNA gene sequences for all the selected species from the respective genomes from the NCBI genome database for the purpose of making species tree.

2.10 Sequence analysis

Pairwise Sequence Alignment is a method for comparing two sequences (DNA, RNA or Protein). This gives us regions of similarity between the two sequences which is helpful in identifying the functional, structural or evolutionary relationships. We used the EMBOSS Needle programme (Rice et al. 2000) for the global pairwise alignment of the protein homologues. This program is based on the Needleman-Wunsch algorithm (Needleman and Wunsch 1970).

Multiple sequence alignments were performed with Clustal Omega (Sievers et al. 2011) with default parameters. Clustal Omega is a program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences.

CD (Conserved Domain) search tool of NCBI (Marchler-Bauer and Bryant 2004) was used to identify the domains in the homologous sequences.

For possible new motif detection, Multiple Expectation Minimization for Motif Elicitation (MEME) program was used (Bailey et al. 2006). MEME is a tool for discovering motifs in a group of related sequences. MEME represents motifs as letter-probability matrices which are position-dependent. Gaps are not incorporated during the motif identification. Patterns having gaps are divided into two or more motifs. It chooses the best motif based on the statistical modelling techniques which depend on the width and number of occurrences of each motif.

The conservation of amino acids within the protein sequence was analyzed by sequence logos. Weblogos 3.2 (Crooks et al. 2004) was used to generate sequence logos. Multiple sequence alignments of DNA or protein sequences can be represented in terms of logos of nucleic acid or amino acids. Every stack in the logo represents the corresponding position in multiple sequence alignments. Sequence conservation can be seen by the height of the stack while the height of each nucleic acid or amino acid represents their frequency in DNA or protein.

2.11 Phylogenetic tree construction

A phylogenetic tree is the pictorial representation of the relationships between organisms. There are several algorithms for determining this relationship. Overall these methods are divided into two categories based on their basic algorithm. First one is distance-based methods which use the amount of dissimilarity (distance) between the aligned sequences to draw trees. Number of differences is called as evolutionary distance. Several algorithms are available in distance-based methods like UPGMA (Unweighted pair group method with Arithmetic mean) (Sokal and Michener 1958), NJ (Neighbor-joining) (Saitou and Nei 1987), FM (Fisch-Margoliash) (Fitch and Margoliash 1967) and ME (Minimum evolution). The second one is the character-based methods which are based directly on the sequence characters rather than a pairwise distance. Two methods fall in this category which is maximum parsimony (MP) and maximum Likelihood (ML). For this study we have used NJ, ML and MP methods.

2.11.1 Neighbor-joining (NJ)

NJ method (Saitou and Nei 1987) is the first choice in the distance-based methods because of its fast computing. It also works when different lineages vary in their rate of evolution. It starts by calculating the evolutionary distances between the sequences based on the evolutionary models and then making a matrix of those distances. At last, it produces a tree based on the distance matrix.

2.11.2 Maximum Likelihood (ML)

Maximum likelihood method is an important method for inferring the evolutionary relationships when the sequences in the study are highly divergent, and their variance is high. This method considers the residues of all the sequences at each site and the log likelihood of these bases are calculated for a given topology by using a probabilistic model. This log-likelihood is added for all the sites, and the sum of the log likelihoods is maximized to estimate the branch length of the tree. This procedure is repeated for all the possible topologies and the topology that shows the highest likelihood is chosen as the final tree.

2.11.3 Maximum Parsimony (MP)

MP (Farris 1970, Fitch 1971) is a simple method used to infer a phylogenetic tree for a set of taxa on the basis of some conserved data on the similarities and differences among taxa. MP method searched for a tree that requires the smallest number of evolutionary changes to explain the differences observed among different Operational taxonomic units (OTU).

2.12 Tree evaluation (Bootstrapping)

Often, any method for tree construction (NJ, ML or MP) is followed by another method called as bootstrapping. Bootstrapping is a statistical technique that tests the sampling errors of a phylogenetic tree by repeatedly sampling trees through slightly changed datasets. The robustness of the original tree can be accessed by this way. In the end, a consensus tree is made which represent the results from all the changed datasets (Soltis and Soltis 2003). Bootstrap gives us an idea about the parts of the tree which are strongly supported with the given data. Normally a 70% bootstrap value represents strong support (Zharkikh and Li 1992).

For our study, we used both Maximum likelihood and Neighbor-Joining methods for the construction of phylogenetic trees. We used MEGA 7.0 (Kumar et al. 2016) for tree construction. Bootstrapping was also performed with 1000 bootstrapping samplings of the sequence data (Felsenstein 1985). We observed that the topologies of both ML and NJ trees are quite similar and the position of clades in the two trees was similar and hence only NJ trees have been discussed in the further analyses.

In spite of taking care of all the necessary details like taking only full sequences, removing the gaps and mismatched regions and trying different substitution models and also different tree construction methods, a large number of nodes in our constructed phylogenetic tree is giving low bootstrap support values. This might happen due to the highly conserved nature of the cyanobacterial species. As cyanobacteria are a unique photosynthetic prokaryote, it might be possible that its genome is highly conserved and hence cannot tell much on its relative evolution within the cyanobacterial class. Again, this could be possible either because there was a common ancestor from which all the genes of these proteins evolved and later become phylogenetically distinct or due to horizontal gene transfer which is quite common in cyanobacteria (Raymond et al. 2002, Rocap et al. 2003, Zhaxybayeva et al. 2006).

2.13 Evolutionary distance calculation

Tamura-Nei (Tamura et al. 2004) method was used for the calculation of evolutionary distances in the gene tree while Jones Taylor Thronton (JTT) method (Jones et al. 1992) was used in case of protein tree.

2.14 Gene duplication and speciation events

For possible gene duplication and speciation events among cyanobacteria, the algorithm described by (Zmasek and Eddy 2001) was used in MEGA 7.0. This algorithm infers speciation and duplication events on a gene tree by comparison to a trusted species tree.

2.15 Protein structure prediction

All the structures in this study were predicted using the homology modelling method. In homology modelling, the protein sequence that is to be modeled (target) shares some similarity with an already known experimentally determined structure (template). The target and template sequences are aligned and then based on the structural information of the template; the target is modeled.

Nitrate assimilation pathway proteins tertiary structures were models for selected homologs. Template search for modelling of these proteins was done by taking the target protein sequence as a query in the BLASTP program and searching this query against the Protein Databank using default parameters. The obtained results were screened for high query coverage and high sequence similarity, and finally, a template was selected.

For our model generation of nitrogen assimilation proteins, we used the standalone version of Modeler 9v15 (Fiser and Sali 2003). This is a program based on the satisfaction of spatial restraints. The model was generated using a single template. We modelled the protein from four nitrogen assimilation pathways proteins. From each order, two representative homologs were modelled based on their respective positions in the phylogenetic tree.

A total of 1000 models were generated through modeller, and the best model was selected on the basis of normalized discrete optimized protein energy score (N-DOPE). The selected best model was energy minimized in GROMACS using the GROMOS96 53a6 force field using steepest descent minimization Algorithms (Van der Spoel et al. 2005).

The quality of a modelled structure is accessed by various methods. These methods use different strategies for the quality assessment. For example, some programs check the stereochemical properties of the model like Ramachandran plot (Ramachandran et al. 1963), PROCHECK (Laskowski et al. 1993), and WHAT-CHECK (Hooft et al. 1996). We have used Ramachandran plot which calculates the overall stereochemical property of the energy minimized model. WHAT-CHECK program was also used to check the protein residue-by-residue and assesses many of its stereochemical properties.

We have also used Verify3d (Eisenberg et al. 1997) which uses a 3D profile to find the relationship of an atomic protein model with its own amino acid sequence. VERIFY3D process by assigning a structural class based on the location and environment of each residue position and by comparing the results to good structures. Environments of residues correspond to three parameters: the local secondary structure, the area of the residue that is buried and the fraction of side-chain area covered by polar atoms.

We also used ERRAT (Colovos and Yeates 1993) which analyzes the non-bonded interactions between the atoms and plots the error function with respect to the position. Errat comparison includes statistics from highly refined structures.

The quality of the models was also evaluated using Qmean Z-score (Benkert et al. 2011) and Qmean score (Benkert et al. 2008) available at Qmean server (Benkert et al. 2009). These scores evaluate the deviations of the predicted model from the crystal structure. Qmean score took into account six parameters (Pairwise, Torsion, All-atom, Solvation, ACC agree and SSE agree) and based on the total score of these parameters a global score ranging from 0 to 1 is given to predict the model reliability. A score near 1 predicts a good model. QMEAN Z-score compares a model with its reference crystal structure and provides the quality of the model. A Z-score less than one is considered as a good quality model, while a score between 1 and 2 and score above two are considered as medium and bad quality models respectively.

Structural analysis, as well as figures, were generated by Visual molecular dynamics (VMD) version 1.9.2 (Humphrey et al. 1996). VMD is a molecular visualization program for displaying, animating, and analyzing large biomolecular systems using 3-D graphics and built-in scripting.

2.16 Identification of functionally important residues in modeled homologs

Arthrospira platensis NIES-39 protein sequence was used as a query in BLAST against the PDB database (Berman et al. 2000) to find the nearest available 3D structure. The top hit with the least E-value was used as the reference. Important residues already identified in the PDB structure were taken from the selected hit, and pairwise sequence alignment was done between the selected hit and the query protein sequence to identify the corresponding important residues in query protein. The identified important residues were compared with all the cyanobacteria species within the MSA. Due to the high identity/similarity of the sequences, a 90% cut off value was set to distinguish conserved and the variable positions.

Chapter III

Functional annotation of the hypothetical proteins of *Arthrospira platensis* NIES-39 genome

3.1 Introduction

Recent advances in high-throughput sequencing techniques like Next Generation Sequencing has led researchers to sequence more genomes. These sequencing projects yield large sequence data for various organisms, which become a part of multiple sequence databases. However, these sequence data are of no use unless they are associated with a function and hence providing a meaningful function to these sequences is a major challenge. Despite all the scientific efforts only about 50-60% of sequences have been annotated in most of the organisms (Goffeau et al. 1996). As most of the cell machinery depends on the proteins for the normal functioning to associate these proteins with proper functions and to understand that how these proteins function in making up a living cell will help the researchers in solving the various aspects of cell functioning.

A genomic annotation normally provides three types of genes, i.e. (a) gene which is functionally annotated (b) hypothetical genes conserved in several organisms and (c) hypothetical genes specific to a genome. All these hypothetical genes give rise to hypothetical proteins (HP) which are thought to be present inside the cell; however, no supporting experimental evidence is available. Results show that these conserved hypothetical proteins were encoded by a substantial fraction of a genome (Galperin and Koonin 2004, Brenchley et al. 2012). These hypothetical proteins may be used as biomarkers as well as other essential signalling proteins viz. Biotic/Abiotic stress proteins (Zarembinski et al. 1998, Doerks et al. 2004). To get insights into the importance of these poorly characterised hypothetical gene/proteins in various physiological developments and stress tolerance issues, it is necessary to annotate these sequences.

There are a number of *in-silico* as well as experimental techniques available for the annotation of the gene sequences and to find meaningful insights into the functional aspects of the identified genes. However, the functional annotation through laboratory experiments would be time consuming and expensive. Hence, bioinformatics tools are the major choice for large-scale functional annotation (Desler et al. 2009). *In silico* methods provide fast and quite reliable results; however, most of these annotation methods are based on the presence of the previously identified sequences. These methods focus on sequence similarity, co-expression, interactions, protein structures etc. (Luo et al. 2007, Horan et al. 2008, Doerks et al. 2012, Schuller et al. 2012). Based on the results of the above methods it assigns a particular function to a query. Since the methods are based mainly on homology, any query which does not give any significant results against the database has to remain un-annotated.

To annotate these un-annotated sequences, we can try the laboratory methods, or we can reuse the *in-silico* methods after some time to see whether some homologous sequence/structure had been made available during that time or not.

The genome of *Arthrospira platensis* NIES-39 had been sequenced and annotated in 2010 (Fujisawa et al. 2010). A total of 6630 protein-coding genes along with 49 RNA genes and 40 tRNA genes were identified. However, Fujisawa et al. analysis could only be able to annotate 5157 (78%) genes while the remaining 1473 (22%) were still un-annotated. It is to be noted that the 78% of annotated genes also included the protein sequences which were homologous to other hypothetical proteins. So, the number of protein sequences with no functional verification is more than 1473. The current total number of genes present in *Arthrospira platensis* NIES-39 is 6666 (NCBI). Out of these, 2622 are hypothetical proteins. In the study, we have tried to functionally annotate the hypothetical protein sequences of *Arthrospira platensis* NIES-39.

3.2 Materials and methods

The general method for annotation starts with the searching of homologous sequences for the hypothetical proteins. Homologous sequences give us an idea about the probable function. The next step is the physiochemical characterisation of protein sequences. Then functional domains of proteins were identified which again provides an idea about the putative protein function. Protein interactions were identified using STRINGS database while pathways in which the protein might be involved are identified using the KEGG database.

3.2.1 Selection of hypothetical proteins for annotation

The complete list of proteins downloaded from genome database has 5872 proteins. Out of these 5872 proteins, 2622 were hypothetical proteins. These 2622 hypothetical proteins were compared against the UniProt database, and finally, 1364 hypothetical genes were selected for annotation (Table 3.1). Remaining genes were either duplicates or present in Uniparc (Uniprot archive).

Table 3.1 List of UniProt Id of hypothetical proteins which were considered for the annotation process.

D4ZMM5	D4ZPN4	D4ZRZ9	D4ZUK2	D4ZWJ0	D4ZYY8	D5A1B5	D5A3V0
D4ZMM6	D4ZPN8	D4ZS00	D4ZUL3	D4ZWJ3	D4ZYY9	D5A1B7	D5A3V1
D4ZMN0	D4ZPN9	D4ZS08	D4ZUM6	D4ZWL1	D4ZYZ3	D5A1B8	D5A3V2
D4ZMP3	D4ZPQ9	D4ZS16	D4ZUM9	D4ZWM4	D4ZYZ5	D5A1C0	D5A3V5
D4ZMP6	D4ZPR6	D4ZS27	D4ZUN1	D4ZWM7	D4ZYZ7	D5A1C3	D5A3W0
D4ZMP7	D4ZPS1	D4ZS31	D4ZUN2	D4ZWN0	D4ZZ02	D5A1C4	D5A3W1
D4ZMQ0	D4ZPS2	D4ZS32	D4ZUN6	D4ZWN4	D4ZZ03	D5A1C8	D5A3X6
D4ZMQ6	D4ZPS6	D4ZS35	D4ZUN8	D4ZWN5	D4ZZ05	D5A1D2	D5A3X7
D4ZMR4	D4ZPS7	D4ZS38	D4ZUP2	D4ZWR2	D4ZZ06	D5A1D4	D5A3Y0
D4ZMS4	D4ZPS9	D4ZS39	D4ZUP8	D4ZWR4	D4ZZ07	D5A1E0	D5A3Y1
D4ZMS5	D4ZPT8	D4ZS40	D4ZUP9	D4ZWS9	D4ZZ08	D5A1E3	D5A3Y2
D4ZMS9	D4ZPT9	D4ZS41	D4ZUQ4	D4ZWT5	D4ZZ19	D5A1E4	D5A3Z2
D4ZMT0	D4ZPU1	D4ZS42	D4ZUQ9	D4ZWU1	D4ZZ21	D5A1E7	D5A3Z3
D4ZMT3	D4ZPU4	D4ZS43	D4ZUR0	D4ZWU2	D4ZZ25	D5A1E8	D5A3Z4
D4ZMT5	D4ZPU6	D4ZS44	D4ZUS2	D4ZWU3	D4ZZ32	D5A1F1	D5A404
D4ZMT7	D4ZPU8	D4ZS45	D4ZUS3	D4ZWU4	D4ZZ34	D5A1G9	D5A418
D4ZMT9	D4ZPU9	D4ZS60	D4ZUS8	D4ZWU5	D4ZZ35	D5A1H5	D5A422
D4ZMU0	D4ZPV0	D4ZS87	D4ZUS9	D4ZWU6	D4ZZ37	D5A1H6	D5A428
D4ZMU3	D4ZPV4	D4ZSA6	D4ZUU3	D4ZWU7	D4ZZ38	D5A1H7	D5A430
D4ZMU4	D4ZPY0	D4ZSA7	D4ZUU4	D4ZWU8	D4ZZ39	D5A1K5	D5A431
D4ZMU6	D4ZPY6	D4ZSB5	D4ZUU7	D4ZWV0	D4ZZ58	D5A1K8	D5A434
D4ZMU8	D4ZPY7	D4ZSB7	D4ZUV0	D4ZWV1	D4ZZ59	D5A1K9	D5A441
D4ZMU9	D4ZPZ0	D4ZSB9	D4ZUV1	D4ZWV8	D4ZZ64	D5A1L2	D5A450
D4ZMV2	D4ZPZ1	D4ZSC0	D4ZUW0	D4ZWV9	D4ZZ65	D5A1M4	D5A460
D4ZMV3	D4ZPZ2	D4ZSC1	D4ZUW3	D4ZWW3	D4ZZ79	D5A1M8	D5A463
D4ZMV8	D4ZPZ4	D4ZSC6	D4ZUW7	D4ZWW5	D4ZZ82	D5A1P1	D5A478
D4ZMW8	D4ZPZ5	D4ZSD4	D4ZUX2	D4ZWY2	D4ZZ87	D5A1P5	D5A483
D4ZMW9	D4ZPZ8	D4ZSD5	D4ZUY7	D4ZWZ0	D4ZZ88	D5A1P8	D5A484
D4ZMX2	D4ZQ03	D4ZSD6	D4ZUY8	D4ZWZ8	D4ZZ90	D5A1Q3	D5A486
D4ZMX3	D4ZQ05	D4ZSD8	D4ZUZ0	D4ZX01	D4ZZB1	D5A1Q6	D5A489
D4ZMX5	D4ZQ07	D4ZSD9	D4ZUZ1	D4ZX08	D4ZZB3	D5A1R4	D5A496
D4ZMX9	D4ZQ11	D4ZSE6	D4ZUZ3	D4ZX10	D4ZZB7	D5A1R7	D5A499
D4ZMY7	D4ZQ13	D4ZSE8	D4ZUZ4	D4ZX15	D4ZZB8	D5A1S3	D5A4A0
D4ZMY9	D4ZQ25	D4ZSE9	D4ZUZ6	D4ZX21	D4ZZB9	D5A1S5	D5A4A1
D4ZMZ0	D4ZQ38	D4ZSF0	D4ZUZ7	D4ZX22	D4ZZC3	D5A1S7	D5A4A2
D4ZMZ5	D4ZQ40	D4ZSF1	D4ZUZ8	D4ZX33	D4ZZC5	D5A1T0	D5A4A3
D4ZMZ9	D4ZQ42	D4ZSF2	D4ZV00	D4ZX39	D4ZZD5	D5A1T4	D5A4B2
D4ZN01	D4ZQ43	D4ZSF6	D4ZV01	D4ZX42	D4ZZD6	D5A1T8	D5A4B4
D4ZN05	D4ZQ44	D4ZSF9	D4ZV07	D4ZX43	D4ZZD8	D5A1U0	D5A4B6
D4ZN07	D4ZQ45	D4ZSG4	D4ZV12	D4ZX47	D4ZZE2	D5A1U5	D5A4B7
D4ZN14	D4ZQ46	D4ZSH6	D4ZV13	D4ZX59	D4ZZE5	D5A1W4	D5A4B9
D4ZN16	D4ZQ47	D4ZSH8	D4ZV24	D4ZX60	D4ZZE6	D5A1W5	D5A4C1
D4ZN21	D4ZQ57	D4ZSH9	D4ZV25	D4ZX81	D4ZZF0	D5A1W8	D5A4C6
D4ZN29	D4ZQ58	D4ZSJ8	D4ZV31	D4ZX90	D4ZZF1	D5A1X0	D5A4D3
D4ZN52	D4ZQ59	D4ZSK3	D4ZV38	D4ZXA0	D4ZZF6	D5A1X3	D5A4E7
D4ZN54	D4ZQ64	D4ZSL3	D4ZV41	D4ZXA3	D4ZZG0	D5A1X4	D5A4F0
D4ZN57	D4ZQ65	D4ZSL4	D4ZV42	D4ZXA7	D4ZZH0	D5A1X7	D5A4F7
D4ZN72	D4ZQ68	D4ZSL6	D4ZV44	D4ZXB8	D4ZZH9	D5A1X9	D5A4G5
D4ZN73	D4ZQ70	D4ZSL7	D4ZV46	D4ZXC5	D4ZZJ0	D5A208	D5A4G9
D4ZN75	D4ZQ73	D4ZSL8	D4ZV47	D4ZXD0	D4ZZJ3	D5A209	D5A4H8
D4ZN76	D4ZQ84	D4ZSP5	D4ZV48	D4ZXD1	D4ZZJ5	D5A215	D5A4H9
D4ZN81	D4ZQ91	D4ZSP8	D4ZV49	D4ZXF5	D4ZZJ6	D5A220	D5A4J9
D4ZN82	D4ZQ99	D4ZSQ0	D4ZV52	D4ZXG0	D4ZZJ7	D5A221	D5A4K9
D4ZN87	D4ZQB6	D4ZSQ1	D4ZV54	D4ZXG1	D4ZZJ8	D5A223	D5A4L0
D4ZNA7	D4ZQC2	D4ZSQ5	D4ZV55	D4ZXG5	D4ZZJ9	D5A229	D5A4L6
D4ZNB1	D4ZQC7	D4ZSQ6	D4ZV56	D4ZXH2	D4ZZK0	D5A231	D5A4L8
D4ZNB5	D4ZQC9	D4ZSR4	D4ZV57	D4ZXI2	D4ZZK4	D5A233	D5A4M2
D4ZNC3	D4ZQD1	D4ZSR5	D4ZV59	D4ZXI3	D4ZZL5	D5A237	D5A4M5
D4ZNC5	D4ZQE5	D4ZST1	D4ZV60	D4ZXK7	D4ZZL7	D5A259	D5A4N3

D4ZNC6	D4ZQE6	D4ZST2	D4ZV61	D4ZXL0	D4ZZM3	D5A260	D5A4N6
D4ZNC7	D4ZQE7	D4ZST6	D4ZV62	D4ZXL1	D4ZZN1	D5A268	D5A4P3
D4ZNC9	D4ZQF5	D4ZST9	D4ZV63	D4ZXL4	D4ZZN4	D5A269	D5A4Q2
D4ZND1	D4ZQF9	D4ZSU0	D4ZV64	D4ZXL8	D4ZZP3	D5A273	D5A4Q3
D4ZND5	D4ZQG0	D4ZSU2	D4ZV65	D4ZXN4	D4ZZQ6	D5A280	D5A4R0
D4ZNE1	D4ZQG3	D4ZSU3	D4ZV66	D4ZXN7	D4ZZS1	D5A293	D5A4S2
D4ZNE3	D4ZQG8	D4ZSV0	D4ZV68	D4ZXN8	D4ZZS3	D5A294	D5A4S6
D4ZNE6	D4ZQG9	D4ZSW0	D4ZV74	D4ZXP2	D4ZZS6	D5A296	D5A4S8
D4ZNF4	D4ZQH1	D4ZSW5	D4ZV75	D4ZXP3	D4ZZT3	D5A298	D5A4S9
D4ZNF7	D4ZQH6	D4ZSW7	D4ZV77	D4ZXR0	D4ZZT5	D5A2A2	D5A4T2
D4ZNG1	D4ZQI2	D4ZSW8	D4ZV80	D4ZXR2	D4ZZT8	D5A2A4	D5A4W4
D4ZNG3	D4ZQI4	D4ZSY1	D4ZV81	D4ZXR3	D4ZZU6	D5A2B2	D5A4W6
D4ZNH1	D4ZQI6	D4ZSZ0	D4ZV85	D4ZXR6	D4ZZU9	D5A2B8	D5A4Y4
D4ZNI0	D4ZQI7	D4ZT18	D4ZV89	D4ZXS0	D4ZZV1	D5A2C2	D5A4Y6
D4ZNI2	D4ZQI8	D4ZT20	D4ZV90	D4ZXS2	D4ZZV2	D5A2D3	D5A4Z0
D4ZNJ2	D4ZQJ1	D4ZT23	D4ZVA3	D4ZXS3	D4ZZV5	D5A2D5	D5A4Z1
D4ZNJ3	D4ZQJ7	D4ZT30	D4ZVA7	D4ZXS5	D4ZZV6	D5A2D6	D5A500
D4ZNK1	D4ZQJ8	D4ZT54	D4ZVA9	D4ZXT0	D4ZZV7	D5A2D7	D5A518
D4ZNK4	D4ZQK1	D4ZT55	D4ZVB0	D4ZXT9	D4ZZW0	D5A2D8	D5A530
D4ZNK5	D4ZQK3	D4ZT65	D4ZVB3	D4ZXU5	D4ZZX0	D5A2D9	D5A541
D4ZNK6	D4ZQK8	D4ZT86	D4ZVB7	D4ZXV8	D4ZZX3	D5A2E3	D5A543
D4ZNK7	D4ZQK9	D4ZT87	D4ZVB9	D4ZXW7	D4ZZZ7	D5A2F0	D5A550
D4ZNL2	D4ZQL0	D4ZT90	D4ZVC0	D4ZXW8	D5A003	D5A2F1	D5A552
D4ZNL3	D4ZQL1	D4ZT96	D4ZVC4	D4ZXY1	D5A023	D5A2F2	D5A561
D4ZNL6	D4ZQL2	D4ZTC1	D4ZVC5	D4ZXY2	D5A024	D5A2F8	D5A573
D4ZNP3	D4ZQL4	D4ZTC2	D4ZVC7	D4ZXZ3	D5A025	D5A2F9	D5A584
D4ZNP5	D4ZQL5	D4ZTD2	D4ZVC8	D4ZXZ4	D5A038	D5A2G8	D5A588
D4ZNP9	D4ZQL6	D4ZTD7	D4ZVC9	D4ZXZ5	D5A039	D5A2J8	D5A592
D4ZNQ2	D4ZQL8	D4ZTE6	D4ZVD2	D4ZXZ9	D5A040	D5A2M6	D5A593
D4ZNQ3	D4ZQN0	D4ZTE7	D4ZVD3	D4ZY04	D5A041	D5A2M9	D5A597
D4ZNQ5	D4ZQP0	D4ZTE9	D4ZVD6	D4ZY05	D5A043	D5A2N1	D5A5A0
D4ZNR8	D4ZQP3	D4ZTG0	D4ZVD7	D4ZY06	D5A044	D5A2Q0	D5A5A1
D4ZNR9	D4ZQP4	D4ZTG7	D4ZVE0	D4ZY12	D5A045	D5A2Q9	D5A5C1
D4ZNS0	D4ZQP6	D4ZTI1	D4ZVE1	D4ZY13	D5A047	D5A2R4	D5A5C2
D4ZNS3	D4ZQP7	D4ZTI2	D4ZVE2	D4ZY17	D5A049	D5A2R8	D5A5C6
D4ZNS6	D4ZQP8	D4ZTI3	D4ZVF0	D4ZY20	D5A063	D5A2S6	D5A5C7
D4ZNS9	D4ZQQ5	D4ZTJ6	D4ZVF1	D4ZY27	D5A074	D5A2S7	D5A5F5
D4ZNT1	D4ZQQ9	D4ZTJ7	D4ZVF7	D4ZY44	D5A076	D5A2T2	D5A5F9
D4ZNT2	D4ZQR2	D4ZTK8	D4ZVH3	D4ZY49	D5A081	D5A2T7	D5A5G1
D4ZNT5	D4ZQR3	D4ZTL0	D4ZVJ0	D4ZY51	D5A086	D5A2U3	D5A5H3
D4ZNU5	D4ZQS5	D4ZTL6	D4ZVM2	D4ZY52	D5A096	D5A2U5	D5A5I2
D4ZNU6	D4ZQS7	D4ZTL7	D4ZVM6	D4ZY58	D5A0A2	D5A2U6	D5A5I6
D4ZNU7	D4ZQT0	D4ZTM1	D4ZVM7	D4ZY60	D5A0A3	D5A2U8	D5A5J0
D4ZNU9	D4ZQT6	D4ZTM9	D4ZVN3	D4ZY61	D5A0B2	D5A2V0	D5A5J1
D4ZNV1	D4ZQT7	D4ZTN2	D4ZVP3	D4ZY63	D5A0B4	D5A2V5	D5A5J3
D4ZNV5	D4ZQU8	D4ZTN5	D4ZVP5	D4ZY64	D5A0B5	D5A2W3	D5A5J6
D4ZNW2	D4ZQV6	D4ZTN8	D4ZVP9	D4ZY70	D5A0C1	D5A300	D5A5J7
D4ZNW5	D4ZQV8	D4ZTP1	D4ZVQ2	D4ZY75	D5A0D3	D5A312	D5A5K0
D4ZNW8	D4ZQV9	D4ZTP4	D4ZVR6	D4ZY76	D5A0D4	D5A316	D5A5K7
D4ZNW9	D4ZQW0	D4ZTQ9	D4ZVS0	D4ZY79	D5A0E5	D5A317	D5A5K8
D4ZNX3	D4ZQX1	D4ZTS6	D4ZVS1	D4ZY83	D5A0E6	D5A318	D5A5K9
D4ZNY3	D4ZQZ9	D4ZTT8	D4ZVS4	D4ZY86	D5A0E8	D5A331	D5A5L4
D4ZNY4	D4ZR14	D4ZTU0	D4ZVS7	D4ZY88	D5A0F0	D5A340	D5A5L5
D4ZNY6	D4ZR24	D4ZTU7	D4ZVT1	D4ZY89	D5A0F4	D5A341	D5A5L6
D4ZNY7	D4ZR26	D4ZTV2	D4ZVT3	D4ZY90	D5A0G2	D5A342	D5A5M1
D4ZNY8	D4ZR35	D4ZTV3	D4ZVT7	D4ZY95	D5A0H6	D5A353	D5A5M5
D4ZNY9	D4ZR49	D4ZTV6	D4ZVU1	D4ZYA0	D5A0H7	D5A362	D5A5M7
D4ZNZ1	D4ZR51	D4ZTV8	D4ZVU2	D4ZYA1	D5A0H9	D5A364	D5A5M8
D4ZNZ2	D4ZR55	D4ZTW6	D4ZVU4	D4ZYA6	D5A0I5	D5A366	D5A5P2
D4ZNZ8	D4ZR76	D4ZTW7	D4ZVU9	D4ZYC3	D5A0I9	D5A370	D5A5R3
D4ZP05	D4ZR99	D4ZTY2	D4ZVV0	D4ZYC5	D5A0J0	D5A373	D5A5T7
D4ZP06	D4ZRA0	D4ZTY3	D4ZVV1	D4ZYD0	D5A0J7	D5A389	D5A5U1
D4ZP07	D4ZRB4	D4ZTZ0	D4ZVV7	D4ZYD2	D5A0K0	D5A394	D5A5U2

D4ZP13	D4ZRC7	D4ZTZ9	D4ZVV9	D4ZYD7	D5A0K2	D5A398	D5A5X4
D4ZP16	D4ZRD0	D4ZU11	D4ZVW0	D4ZYD8	D5A0M0	D5A3A1	D5A5X5
D4ZP17	D4ZRD6	D4ZU12	D4ZVW4	D4ZYE2	D5A0M4	D5A3B5	D5A5Y4
D4ZP18	D4ZRF7	D4ZU17	D4ZVW9	D4ZYF9	D5A0M5	D5A3B7	D5A5Y6
D4ZP26	D4ZRG1	D4ZU27	D4ZVX1	D4ZYG0	D5A0P3	D5A3C5	D5A5Y7
D4ZP28	D4ZRG5	D4ZU36	D4ZVX2	D4ZYG3	D5A0P5	D5A3C6	D5A5Y8
D4ZP31	D4ZRH0	D4ZU38	D4ZVX4	D4ZYG4	D5A0Q0	D5A3C7	D5A5Z3
D4ZP34	D4ZRH1	D4ZU43	D4ZVX5	D4ZYG5	D5A0R1	D5A3C9	D5A5Z7
D4ZP37	D4ZRH4	D4ZU44	D4ZVX9	D4ZYG6	D5A0R2	D5A3D2	D5A5Z9
D4ZP40	D4ZRH5	D4ZU45	D4ZVY0	D4ZYH4	D5A0R8	D5A3D4	D5A600
D4ZP55	D4ZRI1	D4ZU46	D4ZVY2	D4ZYH7	D5A0R9	D5A3D7	D5A601
D4ZP63	D4ZRI4	D4ZU47	D4ZVY6	D4ZYI2	D5A0S2	D5A3D9	D5A604
D4ZP82	D4ZRI5	D4ZU49	D4ZVY7	D4ZYI3	D5A0S3	D5A3E0	D5A608
D4ZP83	D4ZRI7	D4ZU50	D4ZVZ5	D4ZYI8	D5A0S6	D5A3G0	D5A609
D4ZP87	D4ZRJ4	D4ZU53	D4ZVZ8	D4ZYI9	D5A0S7	D5A3H9	D5A632
D4ZP90	D4ZRK1	D4ZU56	D4ZW11	D4ZYJ3	D5A0S9	D5A3I0	D5A652
D4ZP94	D4ZRL1	D4ZU62	D4ZW24	D4ZYJ6	D5A0T2	D5A3I1	D5A654
D4ZP96	D4ZRL2	D4ZU64	D4ZW26	D4ZYJ9	D5A0T6	D5A3I2	D5A656
D4ZP97	D4ZRL5	D4ZU65	D4ZW41	D4ZYK3	D5A0U0	D5A3I3	D5A664
D4ZP98	D4ZRL7	D4ZU67	D4ZW44	D4ZYL3	D5A0W1	D5A3I4	D5A665
D4ZP99	D4ZRL8	D4ZU73	D4ZW45	D4ZYM3	D5A0W3	D5A3I8	D5A679
D4ZPA0	D4ZRL9	D4ZU74	D4ZW47	D4ZYM4	D5A0X2	D5A3I9	D5A680
D4ZPA1	D4ZRM1	D4ZU77	D4ZW48	D4ZYM6	D5A0X3	D5A3J1	D5A681
D4ZPA2	D4ZRN2	D4ZU81	D4ZW52	D4ZYM7	D5A0X6	D5A3J2	D5A682
D4ZPA3	D4ZRN3	D4ZU85	D4ZW61	D4ZYM8	D5A0Y2	D5A3J3	D5A685
D4ZPA4	D4ZRP5	D4ZU86	D4ZW68	D4ZYM9	D5A0Y3	D5A3J8	D5A6A0
D4ZPA6	D4ZRQ9	D4ZU87	D4ZW70	D4ZYN0	D5A0Z5	D5A3L5	D5A6A5
D4ZPA7	D4ZRR7	D4ZU89	D4ZW73	D4ZYN2	D5A107	D5A3L6	D5A6B0
D4ZPB8	D4ZRT1	D4ZU92	D4ZW75	D4ZYN4	D5A118	D5A3L7	D5A6B6
D4ZPC6	D4ZRT6	D4ZU97	D4ZW79	D4ZYN7	D5A121	D5A3M2	D5A6B7
D4ZPC9	D4ZRT7	D4ZUB4	D4ZW84	D4ZYN8	D5A130	D5A3M5	D5A6C2
D4ZPD5	D4ZRT8	D4ZUC2	D4ZW87	D4ZYP0	D5A136	D5A3M6	D5A6C4
D4ZPF1	D4ZRV5	D4ZUC3	D4ZW88	D4ZYP7	D5A144	D5A3M8	D5A6D5
D4ZPF2	D4ZRV7	D4ZUC4	D4ZW99	D4ZYQ3	D5A149	D5A3M9	D5A6F0
D4ZPF3	D4ZRV8	D4ZUC9	D4ZWA4	D4ZYQ5	D5A151	D5A3N9	D5A6F2
D4ZPF5	D4ZRW5	D4ZUD2	D4ZWA8	D4ZYQ6	D5A152	D5A3P3	D5A6F3
D4ZPF6	D4ZRW7	D4ZUD4	D4ZWD4	D4ZYQ7	D5A156	D5A3Q5	D5A6F4
D4ZPF9	D4ZRX6	D4ZUD6	D4ZWD9	D4ZYQ9	D5A158	D5A3Q6	D5A6F9
D4ZPJ6	D4ZRY1	D4ZUD7	D4ZWF3	D4ZYR5	D5A162	D5A3Q7	D5A6G4
D4ZPJ8	D4ZRY2	D4ZUD8	D4ZWF7	D4ZYR8	D5A165	D5A3Q9	D5A6G5
D4ZPJ9	D4ZRY7	D4ZUD9	D4ZWF8	D4ZYR9	D5A170	D5A3S2	D5A6H0
D4ZPK6	D4ZRY9	D4ZUE0	D4ZWG5	D4ZYU7	D5A183	D5A3S4	D5A6I5
D4ZPL0	D4ZRZ0	D4ZUF0	D4ZWH5	D4ZYU8	D5A184	D5A3T2	D5A6I8
D4ZPL2	D4ZRZ3	D4ZUF1	D4ZWH6	D4ZYW4	D5A189	D5A3T5	D5A6J4
D4ZPL7	D4ZRZ4	D4ZUF4	D4ZWH9	D4ZYW7	D5A196	D5A3T9	D5A6K3
D4ZPL8	D4ZRZ5	D4ZUH0	D4ZWI0	D4ZYW9	D5A1A4	D5A3U0	
D4ZPM3	D4ZRZ6	D4ZUI2	D4ZWI2	D4ZYX7	D5A1A7	D5A3U1	
D4ZPM8	D4ZRZ7	D4ZUI9	D4ZWI3	D4ZYY0	D5A1B2	D5A3U3	
D4ZPN3	D4ZRZ8	D4ZUJ2	D4ZWI4	D4ZYY6	D5A1B3	D5A3U9	

3.3 Results and Discussions

3.3.1 Functional annotation using homolog searching

As mentioned, all the selected hypothetical protein sequences were searched using BLASTp for any annotated homologous sequences. After successful searching, we were able to annotate the sequences from 526 hypothetical proteins (Table 3.2). These annotated proteins can be categorised into ten different groups which are enzymes, reverse transcriptase, membrane proteins, endonuclease, recombinase, transcriptional regulators, biosynthetic reactions, nucleic acid binding proteins, ATP binding proteins and other proteins (Figure 3.1). Each of these groups has been discussed here.

Table 3.2 List of the hypothetical proteins along with their annotated functions and physicochemical properties. Functions were allocated by searching the closest homolog of known function. pI, Molecular weight and GRAVY index are mentioned for all the annotated proteins.

UniProt ID	Protein ID	Annotated Function	Protein pI	Molecular Weight	GRAVY
D4ZW68	WP_014273899.1	aldolase	5.58	60186.85	-0.105
D4ZUE0	WP_006619028.1	ATP binding protein	5.03	30873.15	-0.43
D4ZX01	WP_014276322.1	ATP binding protein	5.27	132089.08	-0.321
D4ZR99	WP_006618313.1	ATP binding protein	5.21	52541	-0.184
D4ZN72	WP_006616250.1	ATP binding protein	5.59	51251.39	-0.339
D5A1U0	WP_006618697.1	AAA family ATPase	7.87	8171.42	-0.175
D5A183	WP_006617671.1	AAA family ATPase	5.32	52770.5	-0.304
D4ZQ05	WP_006620232.1	ATPase	6.54	50971.94	-0.199
D5A5I6	WP_014274952.1	ATPase	5.05	33684.53	0.128
D4ZQE7	WP_014277486.1	ATPase	5.72	56539.85	-0.208
D5A2D8	WP_014274593.1	cell division protein ATPase	5.09	122926.41	-0.416
D5A5G1	WP_006618299.1	biotin carboxylase	6.24	15862.19	0.05
D4ZYU7	WP_006618552.1	cobyrinic acid a,c-diamide synthase	6.54	39179.16	-0.013
D5A1F1	WP_006616956.1	cobyrinic acid a,c-diamide synthase	6.61	32229.15	-0.293
D4ZWF7	WP_014273963.1	CocE/NonD hydrolase	4.6	62308.91	-0.237
D5A652	WP_014277147.1	glutamate biogenesis protein	8.89	67588.27	-0.192
D4ZWY2	WP_006617559.1	LPS biosynthesis protein	8.32	31242.79	-0.317
D4ZVV7	WP_014276228.1	LPS biosynthesis protein	5.06	82756.37	-0.333
D4ZU97	WP_014275889.1	oxoacyl ACP synthase	4.65	10646.74	-0.946
D4ZN21	WP_014275099.1	sufE family	8.46	15777.08	-0.369
D4ZWM4	WP_006618852.1	amino oxononanoate synthase	4.49	8749.99	-0.345
D4ZXT0	WP_006616153.1	arginyl tRNA synthetase	9.18	15210.55	0.009
D4ZRG1	WP_006617081.1	arginyl tRNA synthetase	5.63	31363.6	0.155
D5A5U2	WP_006616576.1	ATP synthase	5.38	25366.87	-0.237
D4ZNW9	WP_006617453.1	cyanobactin biosynthesis	4.47	36888.2	-0.165
D5A4L0	WP_006620025.1	cytochrome c biogenesis	9.51	11116.21	0.739
D5A0H7	WP_014276679.1	dethiobiotin synthase	6.56	12982.11	-0.118
D4ZTN2	WP_006620011.1	glucosyl 3 phosphoglycerate synthase	5.66	48323.34	-0.278
D4ZYW7	WP_006618052.1	isochorismatase synthase	6.04	5507.47	-0.188
D4ZN52	WP_014275114.1	lipid a disaccharide synthetase	7.6	44626.59	-0.048
D4ZZ79	WP_014274223.1	lipid a disaccharide synthetase	8.57	47404.05	0.036
D5A317	WP_014276872.1	methionine synthase	6.26	34627.53	-0.399
D5A2N1	WP_006619296.1	methionine synthase	5.37	38605.67	-0.387
D4ZP37	WP_014277373.1	calcium binding protein	4.17	14904.42	-0.325
D4ZVU4	WP_006618202.1	calcium binding protein	4.63	25960.07	-0.645
D4ZUB4	WP_014275901.1	calcium binding protein	6.16	34042.9	-0.299
D5A1R7	WP_014276765.1	calcium binding protein	4.04	33560.03	-0.21
D4ZPL8	WP_014275244.1	calcium binding protein	4.34	59648.16	-0.503
D4ZXA0	WP_006619402.1	chromosome partitioning protein ParB	6.86	28735.86	-0.485
D4ZPN8	WP_014275256.1	chromosome segregation protein	6	8939.26	-0.145
D4ZZH9	WP_014274296.1	chromosome segregation protein	7.78	50545.88	-0.411
D4ZSE8	WP_006616607.1	spor domain protein	9.83	6993.02	-0.498
D4ZSF6	WP_006617692.1	cmr4 protein	4.84	14839.79	-0.43
D4ZUC3	WP_014275906.1	CRISPR associated protein	8.65	90159.7	-0.515
D4ZSF9	WP_006617689.1	CRISPR associated protein	5.72	43884.1	-0.415
D5A1S7	WP_014276770.1	CRISPR associated protein	8.46	44526.97	-0.483
D4ZUC2	WP_006616732.1	CRISPR associated protein	5.68	56284.35	-0.216
D4ZX59	WP_006618778.1	SaqB/ThcOx family dehydrogenase	5.68	57935.32	-0.363
D5A1X0	WP_006619919.1	sterol desaturase	9.52	19157.52	0.385
D4ZTM1	WP_006620002.1	DNA binding protein	5.71	15525.8	0.074
D4ZMR4	WP_014275046.1	DNA binding protein	9.68	75831.24	0.122

D5A3Y0	WP 014274768.1	DNA polymerase	4.97	11200.76	-0.398
D4ZWH6	WP 014273971.1	DNA polymerase III	6.15	98452.34	-0.418
D5A341	WP 006618857.1	DNA repair	5.76	45248.96	-0.661
D5A6F4	WP 014277218.1	primosomal protein	5.45	33775.13	-0.566
D4ZSY1	WP 014275641.1	bstEII	8.27	25483.32	-0.41
D4ZZ34	WP 006617952.1	endoU nuclease	9.67	21226.68	-0.389
D5A0B2	WP 014276648.1	hnh endonuclease	11	6115.24	0.174
D5A3B7	WP 006617505.1	hnh endonuclease	10.45	6506.58	-0.379
D5A6F9	WP 014277222.1	hnh endonuclease	6.01	24234.53	-0.795
D5A1X3	WP 014276798.1	hnh endonuclease	10.51	47317.41	-0.474
D5A2S7	WP 014274670.1	hnh endonuclease	10.49	47470.76	-0.465
D4ZPU9	WP 014275287.1	hnh endonuclease	11.66	8096.43	-1.167
D4ZS08	WP 014275466.1	hnh endonuclease	10.38	48183.16	-0.483
D4ZS27	WP 014275479.1	hnh endonuclease	10.48	48795.83	-0.499
D4ZZ65	WP 014274215.1	hnh endonuclease	11.46	8126.45	-1.118
D5A3U1	WP 006616396.1	restriction endonucleases	5.91	16839.16	-0.182
D4ZPZ8	WP 014277406.1	restriction endonucleases	6.28	45029.65	-0.319
D4ZNW2	WP 006617460.1	restriction endonucleases	6.16	11641.02	0.761
D4ZSF2	WP 006617693.1	restriction endonucleases	4.77	24266.96	0.003
D4ZZN1	WP 014274330.1	restriction endonucleases	5.44	128029.18	-0.421
D5A1L2	WP 014276737.1	restriction endonucleases	5.04	53715.74	-0.284
D5A121	WP 006619157.1	ribonuclease HI	4.77	10355.52	-0.597
D5A118	WP 006619154.1	SnaBI endonuclease	5.96	24772.2	-0.283
D4ZMP7	WP 006616503.1	uma2 family endonuclease	4.62	29938.7	-0.797
D4ZUF4	WP 006617750.1	uma2 family endonuclease	4.51	31145.09	-0.474
D4ZVY6	WP 014276245.1	uma2 family endonuclease	4.74	35077.73	-0.936
D4ZU77	WP 006618015.1	uma2 family endonuclease	5	31478.1	-0.669
D4ZYD8	WP 014276435.1	uma2 family endonuclease	4.94	30324.22	-0.682
D4ZYD7	WP 014276434.1	uma2 family endonuclease	4.82	28541.24	-0.571
D4ZZ59	WP 014274211.1	uma2 family endonuclease	4.82	28888.43	-0.752
D5A5L6	WP 014274975.1	uma2 family endonuclease	4.68	27454.85	-0.674
D5A5L5	WP 014274974.1	uma2 family endonuclease	4.68	27381.8	-0.654
D4ZNK4	WP 014277276.1	uma2 family endonuclease	4.49	32921.77	-0.602
D5A5K8	WP 014274969.1	uma2 family endonuclease	4.66	28224.66	-0.707
D4ZNK6	WP 014277278.1	uma2 family endonuclease	4.54	31409.17	-0.594
D5A086	WP 006615898.1	uma2 family endonuclease	4.51	27679.29	-0.417
D5A130	WP 014274466.1	uma2 family endonuclease	4.96	28606.25	-0.669
D4ZNK1	WP 014277273.1	uma2 family endonuclease	4.7	28314.87	-0.6
D4ZNK5	WP 014277277.1	uma2 family endonuclease	4.62	29133.66	-0.625
D4ZNK7	WP 014277279.1	uma2 family endonuclease	4.69	27626.1	-0.6
D4ZT30	WP 006617803.1	uma2 family endonuclease	4.68	25951.98	-0.817
D4ZTD7	WP 006618570.1	uma2 family endonuclease	5.03	27094.69	-0.522
D4ZVH3	WP 014276157.1	uma2 family endonuclease	4.91	30687.27	-0.957
D4ZZ58	WP 014274210.1	uma2 family endonuclease	4.77	24602.02	-0.409
D5A1B2	WP 014274514.1	uma2 family endonuclease	4.55	31882.69	-0.516
D5A1B3	WP 014274515.1	uma2 family endonuclease	4.6	31144.88	-0.525
D5A220	WP 014274516.1	uma2 family endonuclease	4.54	33602.7	-0.501
D5A4L6	WP 006620031.1	uma2 family endonuclease	4.63	21304.39	-0.217
D5A5K7	WP 014274968.1	uma2 family endonuclease	4.67	30661.2	-0.813
D5A5K9	WP 014274970.1	uma2 family endonuclease	4.72	30675.23	-0.817
D5A5L4	WP 014274973.1	uma2 family endonuclease	4.72	30689.26	-0.817
D4ZZG0	WP 014274282.1	esterase-like activity	4.76	34938.46	-0.203
D4ZX60	WP 006618777.1	3-5 exonuclease	5.92	93527.65	-0.345
D4ZXZ3	WP 014274116.1	dynamin protein	5.14	82767.72	-0.501
D4ZVQ2	WP 006617250.1	GTPase family protein	5.57	72449.4	-0.157
D4ZZC5	WP 006617311.1	dead/deah box helicase	6.82	11509.36	-0.177
D4ZZB8	WP 014274258.1	dead/deah box helicase	9.18	38666.2	-0.239
D4ZZB9	WP 014274259.1	dead/deah box helicase	7.7	18826.26	-0.395
D4ZZB1	WP 014274251.1	dead/deah box helicase	9.36	11604.27	-0.108
D4ZZC3	WP 014274263.1	dead/deah box helicase	6.57	11891.63	-0.388
D4ZZB7	WP 014274257.1	dead/deah box helicase	6.01	6903.76	-0.558
D4ZYR9	WP 014276503.1	DNA helicase	10.36	28461.03	-0.7
D5A269	WP 006618443.1	DNA helicase	6.87	6787.81	-0.031
D4ZQZ9	WP 014275353.1	DNA helicase	5.71	118522.01	-0.515
D4ZNF7	WP 006616820.1	helicase	6	5585.24	-0.643
D4ZNF4	WP 014277254.1	helicase	8.62	21016.21	-0.367
D5A366	WP 014276897.1	helicase	6.08	59327.34	-0.121
D4ZQG9	WP 014277498.1	hydrogenase	5.41	47414.64	-0.373
D4ZNQ5	WP 014277303.1	aspartoacylase family	6.04	43160.03	-0.279
D5A654	WP 006619612.1	glucosamine 6 phosphate deaminase	9.25	7350.36	-1.283
D4ZRL8	WP 006619847.1	glycoside hydrolase	8.3	49211.85	-0.131

D4ZST2	WP_014275605.1	glycoside hydrolase	4.87	46519.42	-0.449
D4ZWF3	WP_014273959.1	glycosyl hydrolase	5.69	56906.24	-0.47
D4ZTU0	WP_014275800.1	glycosyl hydrolase	8.3	100582.21	-0.318
D4ZV25	WP_014276064.1	HAD family hydrolase	5	30244.87	-0.22
D4ZVU2	WP_006618204.1	haloacid dehalogenase-like hydrolase	8.45	55328.29	0.389
D4ZZV7	WP_014276570.1	hydrogenase maturation protease	4.44	17637.61	-0.351
D5A6C4	WP_014277198.1	hydrolase	6.41	29853.72	-0.011
D5A5Z7	WP_006618594.1	inosine-uridine nucleoside N ribohydrolase	4.89	25404.73	-0.383
D4ZSW8	WP_014275631.1	isochorismatase	5.52	37586.67	-0.177
D5A5Y6	WP_006616421.1	nucleotide pyrophosphatase	4.95	12512.19	-0.809
D4ZP07	WP_014277363.1	oxopronilase	5.53	132890.24	-0.147
D4ZS60	WP_006617989.1	phosphohydrolase	5.78	32366.7	-0.352
D4ZP83	WP_014275182.1	phosphotransacetylase	4.9	39140.07	0.128
D4ZNL3	WP_006617922.1	polysaccharide deacetylase	6.84	35330.58	-0.252
D5A3E0	WP_006617124.1	pyrophosphatase	4.7	21004.43	-0.499
D4ZR26	WP_014275375.1	serine hydrolase	5.17	35687.68	-0.183
D5A4N3	WP_006620045.1	zinc dependent hydrolase	6.55	27757.31	0.079
D4ZN57	WP_006616981.1	zn dependent hydrolase	8.62	29287.46	-0.085
D5A215	WP_014276816.1	alpha amylase	4.79	95854.41	-0.401
D4ZTG7	WP_006617379.1	beta glucosidase	4.94	91120.33	-0.356
D4ZNT5	WP_006616735.1	beta lactamase	8.74	48691.18	-0.473
D4ZV07	WP_014276053.1	beta lactamase	8.74	48546.04	-0.45
D5A4A3	WP_014274845.1	chitinase	5.29	14562.12	-0.685
D5A4A2	WP_006616032.1	chitinase	5.67	13046.88	-0.374
D4ZUI2	WP_014275934.1	creatininase family protein	6.35	27097.83	-0.059
D4ZYQ3	WP_014276495.1	hydroxylase	7.75	30849.99	-0.394
D5A518	WP_014277098.1	diguanylate cyclase	5.9	63993.88	-0.214
D4ZZX0	WP_014276576.1	diguanylate cyclase	7.06	51956.05	-0.132
D4ZUW7	WP_006619963.1	acetylglutamate kinase	4.69	7280.41	-0.077
D4ZTT8	WP_014275799.1	diacylglycerol kinase	6.07	35838.19	-0.052
D5A550	WP_006617101.1	2-5 RNA ligase	4.73	30428.48	-0.352
D5A404	WP_006619374.1	carboxylate amine ligase	6.26	61182.29	-0.421
D5A024	WP_014276599.1	propionate CoA ligase	4.73	7480.47	0.107
D4ZXG0	WP_006617219.1	lipase	5.56	43173.02	-0.323
D4ZXG1	WP_006617220.1	lipase	5.42	46824.62	-0.289
D4ZTV6	WP_006619346.1	lipase	9.18	26616.49	-0.394
D4ZPF9	WP_006617597.1	lipase chaperone	4.91	49851.77	-0.148
D4ZQG3	WP_006619284.1	PEP carboxylase	5.42	22874.97	-0.371
D5A1C4	WP_014276692.1	lysogenization protein	8.64	26917.4	0.754
D5A5J0	WP_014274955.1	energy transducer TonB	4.27	55945.16	-0.591
D4ZR35	WP_006618064.1	fasciclin	8.47	19077.41	-0.378
D5A0I5	WP_006619874.1	fasciclin	4.77	21432.99	-0.18
D4ZPS1	WP_014275274.1	fecR family protein	5.81	33415.99	-0.093
D4ZXS0	WP_014274079.1	mechanosensitive ion channel	6.24	54872.16	0.222
D4ZYG3	WP_006615985.1	membrane bound metallopeptidase	6.08	11008.35	-0.738
D5A5Z9	WP_006618593.1	membrane protein	4.51	37012.8	0.51
D4ZPM8	WP_006620127.1	membrane protein	7.83	20212.92	0.402
D5A5M7	WP_006617761.1	membrane protein	4.86	18948.32	-0.587
D4ZTG0	WP_014275726.1	membrane protein	8.01	28378.97	-0.066
D4ZTI3	WP_006618326.1	membrane protein	9.88	27118.35	0.475
D4ZTI2	WP_014275737.1	membrane protein	5.59	53520.28	0.008
D5A231	WP_014274522.1	membrane protein	10.29	11618.33	1.412
D4ZPK6	WP_006617229.1	membrane protein	4.38	22744.49	-0.391
D4ZVP3	WP_006616691.1	membrane protein	9.63	60356.58	0.36
D5A478	WP_014274831.1	membrane protein	4.85	11138.22	-0.408
D5A1X4	WP_006617163.1	membrane protein	9.69	11392.29	0.705
D4ZRC7	WP_014275422.1	membrane protein	6.83	26059.41	0.145
D4ZTD2	WP_006618575.1	membrane protein	8.31	67204.3	0.395
D4ZVP5	WP_006616903.1	membrane protein	8.45	63620.45	0.488
D4ZW24	WP_014276267.1	membrane protein	5.96	30723.51	-0.009
D4ZWT5	WP_006618288.1	membrane protein	5.15	58836.66	0.864
D5A0I9	WP_006619870.1	membrane protein	8.98	17622.35	1.038
D4ZR24	WP_014275373.1	membrane protein	9.79	34753.36	0.655
D4ZV13	WP_006616094.1	membrane protein	8.85	29127.95	0.193
D4ZYR8	WP_006616811.1	membrane protein	4.15	22283.13	-0.051
D4ZRX6	WP_006617195.1	membrane protein	6.06	11343.95	-0.026
D4ZNZ8	WP_014277354.1	membrane protein	8.83	59140.04	0.602
D4ZZF0	WP_014274276.1	membrane protein	5.44	33081.64	-0.152
D4ZQ03	WP_006620234.1	nfeD like protein	4.88	21917.06	0.289
D5A076	WP_006616540.1	ompA family protein	5.06	49525.44	0.039
D4ZTI1	WP_006618324.1	ompA family protein	4.97	29701.42	-0.277

D4ZZW0	WP_006619167.1	periplasmic hydrogenase	5.58	10470.24	-0.084
D5A2V5	WP_014276833.1	potassium channel	8.12	62957.96	0.218
D4ZQI1	WP_006620226.1	tspO/mbr family	11.51	7575.98	0.192
D4ZVT7	WP_014276214.1	YeeE/YedE family protein	9.36	15455.58	0.867
D5A561	WP_014274876.1	metal binding protein	10.98	10929.78	-0.107
D4ZQE6	WP_014277485.1	metal binding protein	4.89	19266.9	-0.347
D4ZUJ2	WP_014275942.1	metal dependent phosphohydrolase	5.67	36255.87	-0.386
D5A156	WP_006617284.1	metallophosphatase	4.62	26283.83	-0.174
D5A5Y8	WP_006616419.1	tellurite resistance	4.66	11159.74	-0.392
D4ZXS5	WP_014274084.1	tellurite resistance	4.6	17281.99	0.174
D5A229	WP_006619942.1	tellurite resistance	4.46	15270.53	0.007
D4ZQ65	WP_014277435.1	DNA adenine methylase	8.99	33322	-0.567
D4ZZ19	WP_006615831.1	twitching motility protein	6.7	8301.57	0.06
D5A450	WP_006620084.1	type IV pilin	5.29	24077.5	-0.342
D4ZPZ4	WP_006620242.1	nitrate reductase associated protein	5.67	17578.12	-0.35
D4ZPZ5	WP_014277403.1	nuclease	5.62	60528.24	-0.304
D4ZSH9	WP_014275569.1	nuclease	5.94	20156.02	-0.41
D4ZWS9	WP_014274026.1	nuclease NurA	5.72	45751.33	-0.206
D4ZVB9	WP_014276119.1	PIN domain protein	4.73	16847.34	0.024
D4ZY49	WP_006619039.1	nucleic acid binding protein	4.52	7420.25	-0.547
D4ZW88	WP_014273916.1	nucleic acid binding protein	4.97	17796.83	0.3
D5A4S2	WP_014277048.1	nucleic acid binding protein	5.67	16836.59	0.267
D4ZVX2	WP_014276235.1	ketol-acid reductoisomerase	4.91	8631.81	-0.204
D4ZWW3	WP_014276297.1	methanol dehydrogenase	4.64	25951.56	-0.102
D4ZRT8	WP_014275437.1	oxidoreductase	9.24	43365.24	-0.122
D4ZVZ5	WP_006618986.1	oxidoreductase	5.49	36383.42	-0.042
D5A152	WP_006617280.1	pryrimidine nucleotide-disulphide oxidoreductase	9.38	46958.38	-0.165
D4ZPU1	WP_006617795.1	vinyl reductase	5.36	25536.01	-0.308
D4ZPT8	WP_014275284.1	vinyl reductase	5.3	26684.12	-0.342
D5A296	WP_014274559.1	hemerthrin	6.58	16224.7	-0.536
D4ZU27	WP_014275848.1	aspartic protease	4.5	35621.79	0.204
D5A0B4	WP_006616914.1	aspartyl protease	5.1	21527.7	0.254
D5A4D3	WP_006615936.1	aspartyl protease	5.76	15484.97	-0.01
D4ZTU7	WP_006619355.1	ATP dependent zinc protease	5.81	24823.61	0.184
D4ZND5	WP_006618928.1	carboxypeptidase	4.41	15807.33	-0.095
D4ZR55	WP_006618749.1	peptidase	5.79	27887.78	-0.271
D4ZR51	WP_014275386.1	peptidase	5.95	98398.56	-0.188
D4ZMZ9	WP_014275086.1	peptidase	4.32	16386.62	0.057
D4ZPL0	WP_014275236.1	peptidase	5.89	135004.15	-0.03
D4ZV00	WP_014276049.1	peptidase	4.09	52510.7	-0.281
D5A3U9	WP_014274743.1	peptidase	4.13	52423.71	-0.254
D4ZPS6	WP_006616026.1	protease	5.06	15314.6	0.165
D4ZUY8	WP_014276038.1	serine protease	10.15	12791.72	-0.246
D5A588	WP_014274896.1	serine protease	4.13	40459.69	-0.224
D5A2Q0	WP_014274654.1	PBS lyase	4.26	68250.24	-0.663
D5A3J3	WP_014276953.1	PBS lyase	5.27	46275.55	-0.559
D4ZVE2	WP_014276134.1	PBS lyase	4.63	17181.74	-0.19
D4ZVA7	WP_006616892.1	photosystem reaction center subunit	4.68	25417.5	-0.444
D4ZNW5	WP_006617457.1	phasin phaA protein	4.9	10698.14	-0.985
D4ZVS4	WP_006618221.1	TPR protein	5.05	7129.21	-0.078
D4ZZV2	WP_014276568.1	TPR protein	9.51	65908.29	-0.41
D4ZZ88	WP_014274232.1	TPR protein	8.98	8028.32	-0.13
D4ZVC4	WP_014276123.1	TPR protein	9.42	8075.33	-0.248
D4ZVC0	WP_006616072.1	TPR protein	4.35	7678.79	0.132
D4ZNC5	WP_014275153.1	TPR protein	10.93	7667.15	0.307
D4ZNC3	WP_014275151.1	TPR protein	10.93	7653.12	0.276
D5A2T2	WP_014274675.1	TPR protein	6.11	8543.8	-0.164
D5A136	WP_014274471.1	TPR protein	5.79	36269.46	-0.162
D4ZVZ8	WP_006618982.1	TPR protein	5.27	19728.25	-0.347
D4ZWJ3	WP_006617411.1	TPR protein	5.45	33561.43	-0.146
D4ZMP6	WP_014275033.1	TPR protein	9.45	90758.79	-0.029
D5A3P3	WP_006617068.1	TPR protein	5.14	16408.63	-0.54
D5A1C3	WP_006618096.1	ATP12 chaperone protein	5.28	15001.93	0.155
D4ZW70	WP_014273901.1	molecular chaperone	9.22	19704.41	-0.762
D4ZW73	WP_014273904.1	molecular chaperone	8.9	19699.37	-0.715
D4ZW75	WP_014273906.1	molecular chaperone	8.9	19676.34	-0.718
D4ZQP4	WP_014275311.1	recombinase	9.4	55485.71	-0.516
D4ZQL4	WP_014277512.1	recombinase	9.23	55309.55	-0.47
D4ZSL8	WP_014275584.1	recombinase	9.21	55098	-0.516
D5A1M4	WP_014276743.1	recombinase	9.28	55421.44	-0.515

D5A5C1	WP_014274910.1	recombinase	9.3	55263.32	-0.493
D4ZUM6	WP_014275968.1	recombinase	9.39	55282.55	-0.464
D5A4P3	WP_014277026.1	recombinase	9.19	55103.14	-0.47
D4ZMY7	WP_014275080.1	recombinase	9.27	55109.03	-0.509
D4ZZL7	WP_014274325.1	recombinase	9.14	55109.35	-0.459
D5A3U3	WP_014274738.1	recombinase	9.13	55107.26	-0.432
D5A6I8	WP_014277235.1	recombinase	9.1	54989.03	-0.467
D5A3J1	WP_014276951.1	recombinase	9.33	54130.11	-0.463
D4ZQP3	WP_014275310.1	recombinase	10.01	8552.69	-0.368
D5A1B7	WP_014276689.1	recombinase	9.45	18931.25	-0.076
D4ZN76	WP_014274428.1	recombinase	9.35	17011.96	-0.073
D4ZQ38	WP_014277424.1	recombinase	9.16	36870.99	-0.612
D5A0X6	WP_014274425.1	recombinase	9.26	36799.68	-0.672
D5A1H5	WP_014276719.1	recombinase	9.24	36877.78	-0.648
D4ZN73	WP_014275130.1	recombinase	9.09	36925.73	-0.672
D4ZSZ0	WP_014275647.1	recombinase	9.18	36897.68	-0.674
D4ZMV8	WP_014275064.1	recombinase	8.2	23236.15	-0.836
D4ZSG4	WP_014275560.1	recombinase	7.67	23095.9	-0.838
D4ZUL3	WP_014275958.1	recombinase	9.38	33941.67	-0.624
D5A5H3	WP_014274941.1	recombinase	5.74	83366.6	-0.119
D4ZNA7	WP_014275138.1	recombinase	9.23	53448.7	-0.647
D5A1B8	WP_014276690.1	recombinase	9.29	32868.3	-0.782
D4ZSP5	WP_014275593.1	recombinase	9.18	53442.58	-0.644
D4ZV57	WP_014276080.1	recombinase	9.33	53453.7	-0.647
D5A023	WP_014276598.1	recombinase	9.33	53467.72	-0.642
D5A4S9	WP_014277052.1	recombinase	9.22	53466.78	-0.657
D4ZUK2	WP_014275949.1	recombinase	8.78	41012.38	-0.646
D4ZYA1	WP_014274184.1	recombinase	9.52	9636.04	-0.65
D4ZPV4	WP_006615945.1	recombinase	5.31	18458.97	-0.137
D5A1R4	WP_014276762.1	recombinase	9.09	53319.57	-0.584
D4ZQL8	WP_014277515.1	recombinase	8.88	65629.54	-0.401
D5A0B5	WP_014276649.1	camA protein	7.94	8975.17	-0.812
D4ZNY4	WP_006618038.1	dehydrogenase/reductase	4.38	8275.96	-1.014
D4ZPB8	WP_014275202.1	calvin cycle regulation	4.04	60056.28	-0.518
D4ZXN4	WP_006616414.1	FHA domain protein	4.3	6681.46	-0.443
D4ZYI3	WP_014276462.1	histone acetylation	11.45	8152.49	-0.982
D4ZTJ7	WP_014275741.1	LtrA protein	11.11	8259.63	-1.001
D5A3V5	WP_014274746.1	LtrA protein	11.36	9036.48	-1.055
D5A0A2	WP_014276641.1	LtrA protein	11.33	8271.61	-1.022
D5A1H7	WP_014276721.1	LtrA protein	11.59	8377.68	-1.241
D4ZMS5	WP_014274625.1	LtrA protein	11.92	8375.76	-1.182
D4ZMW9	WP_014275017.1	LtrA protein	11.81	8248.57	-1.159
D4ZSC0	WP_014275534.1	LtrA protein	11.58	8257.57	-1.103
D4ZU11	WP_014275841.1	LtrA protein	11.66	8257.62	-1.115
D5A3W1	WP_014274751.1	LtrA protein	11.58	8253.54	-1.079
D4ZMX2	WP_014275071.1	LtrA protein	11.58	8249.55	-1.092
D4ZSD5	WP_014275546.1	LtrA protein	11.81	8276.62	-1.126
D4ZTW7	WP_014275811.1	LtrA protein	11.58	8277.56	-1.111
D4ZTZ0	WP_014275826.1	LtrA protein	11.81	8934.34	-1.079
D5A3Z2	WP_014274779.1	LtrA protein	11.66	8234.58	-1.119
D5A3D7	WP_014276931.1	LtrA protein	11.59	8949.31	-1.061
D5A5Z3	WP_014277121.1	LtrA protein	11.39	8866.22	-1.059
D4ZNB1	WP_014275142.1	LtrA protein	11.59	8926.29	-1.095
D4ZVF0	WP_014276140.1	LtrA protein	11.81	8234.54	-1.16
D5A3C7	WP_014276925.1	LtrA protein	9.26	17039.04	-0.254
D4ZQN0	WP_014275301.1	reverse transcriptase	12.02	6724.04	-0.504
D5A3G0	WP_014276937.1	reverse transcriptase	11.49	8211.76	-0.515
D4ZRK1	WP_014274351.1	reverse transcriptase	11.86	16541.51	-0.867
D4ZYI2	WP_014276461.1	reverse transcriptase	10.19	8244.36	-0.634
D4ZQT7	WP_014275336.1	reverse transcriptase	9.89	8323.48	-0.653
D4ZTJ6	WP_014275740.1	reverse transcriptase	9.91	8262.48	-0.482
D4ZVM2	WP_014276175.1	reverse transcriptase	11.6	15843.69	-0.871
D4ZVM6	WP_014276179.1	reverse transcriptase	11.97	15829.65	-0.865
D5A3C5	WP_014276923.1	reverse transcriptase	11.86	15781.65	-0.82
D4ZMX3	WP_014275072.1	reverse transcriptase	9.2	8557.66	-0.632
D4ZUQ4	WP_014275985.1	reverse transcriptase	11.9	8457.29	-0.292
D5A0A3	WP_014276642.1	reverse transcriptase	9.89	8244.4	-0.558
D4ZPV0	WP_014275288.1	reverse transcriptase	9.98	8282.43	-0.674
D5A3D2	WP_014274752.1	reverse transcriptase	9.2	8566.68	-0.628
D4ZMW8	WP_014275016.1	reverse transcriptase	12.03	8527.3	-0.323
D4ZSD6	WP_014275547.1	reverse transcriptase	8.82	8543.68	-0.511
		reverse transcriptase			

D4ZVF1	WP_014276141.1	reverse transcriptase	12.03	8512.25	-0.393
D5A3A1	WP_014276917.1	reverse transcriptase	11.78	8547.24	-0.532
D5A6K3	WP_014277242.1	reverse transcriptase	11.92	8579.25	-0.528
D4ZP40	WP_014276363.1	reverse transcriptase	7.88	8505.59	-0.475
D4ZQ59	WP_014275538.1	reverse transcriptase	12.15	8586.37	-0.488
D4ZSB7	WP_014275532.1	reverse transcriptase	12.03	8585.3	-0.457
D4ZU12	WP_014275842.1	reverse transcriptase	11.92	8545.23	-0.519
D4ZVM7	WP_014276180.1	reverse transcriptase	9.2	8580.7	-0.601
D4ZZB3	WP_014274253.1	reverse transcriptase	12.04	8514.26	-0.445
D5A1Q3	WP_014276756.1	reverse transcriptase	8.85	8553.68	-0.526
D5A3C6	WP_014276924.1	reverse transcriptase	9.98	8192.28	-0.612
D5A3W0	WP_014274750.1	reverse transcriptase	11.91	8557.24	-0.443
D4ZN75	WP_014274427.1	reverse transcriptase	11.78	8577.31	-0.431
D4ZNE1	WP_014277243.1	reverse transcriptase	12.04	8572.38	-0.335
D4ZQ40	WP_014277426.1	reverse transcriptase	11.78	8589.37	-0.361
D4ZR14	WP_014275364.1	reverse transcriptase	11.78	8519.23	-0.43
D4ZSB5	WP_014275530.1	reverse transcriptase	9.2	8495.64	-0.561
D4ZSC1	WP_014275535.1	reverse transcriptase	11.78	8545.32	-0.368
D4ZSD4	WP_014275545.1	reverse transcriptase	11.92	8557.29	-0.449
D4ZTV8	WP_014275805.1	reverse transcriptase	11.92	8589.33	-0.404
D4ZTW6	WP_014275810.1	reverse transcriptase	11.78	8511.26	-0.491
D4ZYJ3	WP_014276468.1	reverse transcriptase	11.78	8500.32	-0.272
D4ZZ64	WP_014274214.1	reverse transcriptase	11.92	8573.29	-0.486
D5A0T2	WP_014274388.1	reverse transcriptase	11.78	8457.25	-0.308
D5A1H6	WP_014276720.1	reverse transcriptase	12.15	8556.34	-0.454
D5A3Z3	WP_014274780.1	reverse transcriptase	11.78	8526.35	-0.308
D4ZMS4	WP_014274626.1	reverse transcriptase	12.15	8529.32	-0.418
D4ZNB5	WP_014275146.1	reverse transcriptase	11.78	8585.38	-0.376
D4ZPU8	WP_014275286.1	reverse transcriptase	11.78	8587.35	-0.442
D4ZQT6	WP_014275335.1	reverse transcriptase	11.58	15637.38	-0.81
D4ZSB9	WP_014274255.1	reverse transcriptase	8.85	8550.63	-0.591
D5A5R3	WP_014275005.1	reverse transcriptase	5.89	11153.69	-0.413
D4ZQE5	WP_014277484.1	RNA binding protein	4.88	17783.92	-0.601
D4ZQS7	WP_006616290.1	RNA binding protein hfq	4.85	8115.49	-0.334
D5A2U6	WP_014276827.1	RNA polymerase	5.73	24370.96	-0.116
D4ZRY9	WP_014274803.1	RNA polymerase	5.83	32490.94	-0.199
D4ZP96	WP_014275190.1	RNA polymerase	5.82	32417.71	-0.195
D4ZNY3	WP_006618037.1	fist N domain protein	4.77	44943.33	-0.046
D5A601	WP_006618591.1	heavy metal sensor	11.26	14738.35	-0.844
D5A2B2	WP_014274572.1	sensory protein	9.74	28926.42	0.836
D5A370	WP_006620273.1	adenylate cyclase	4.27	16815.98	-0.171
D5A4G5	WP_014276985.1	adenylate cyclase	5.54	85142.62	-0.344
D5A293	WP_014274557.1	guanylase cyclase	5.56	55349.02	-0.41
D5A0M4	WP_006619573.1	histidine kinase	4.88	15118.69	0.37
D4ZQL5	WP_006616474.1	histidine kinase	4.75	6801.87	0.344
D5A0M5	WP_006619572.1	histidine kinase	4.55	7147.9	-0.434
D4ZS16	WP_014275472.1	histidine kinase	6.25	7195.25	-0.362
D4ZST1	WP_006619193.1	histidine kinase	5.13	11357.91	-0.59
D5A0W3	WP_014274414.1	signal transduction protein	6.16	8115.13	-0.42
D4ZTY2	WP_006616111.1	alpha crystallin family protein	5.73	13250.41	-1.288
D4ZPF6	WP_006617600.1	nirD stress tolerance protein	6.29	11699.87	-0.292
D4ZP13	WP_006618371.1	s layer domain protein	4.65	17695.09	-0.359
D4ZPS2	WP_006620161.1	s layer domain protein	9.95	22683.06	0.178
D4ZQW0	WP_006616674.1	s layer domain protein	4.7	8554.68	0.172
D5A170	WP_014274495.1	s layer domain protein	4.82	58236.41	-0.421
D4ZMQ0	WP_014275035.1	s layer domain protein	8.74	84687.26	-0.143
D4ZSU2	WP_006619182.1	baseplate protein	5.12	15584.02	-0.188
D4ZYY6	WP_014276549.1	flagellar protein	7.66	18691.53	-0.282
D4ZRH5	WP_006617730.1	phage tail protein	5.13	20131.53	-0.405
D4ZMV2	WP_014275060.1	tail fiber domain protein	9.04	18979.37	-0.315
D4ZMZ0	WP_014275083.1	tail fiber domain protein	5.66	26829.08	-0.326
D4ZSK3	WP_014275575.1	tail fiber domain protein	5.95	26982.26	-0.367
D4ZV77	WP_014276099.1	tail fiber domain protein	5.66	26832.01	-0.366
D4ZSU0	WP_014275612.1	tail protein	4.83	41466.42	-0.109
D4ZPS7	WP_006616027.1	toxin antitoxin system	4.28	9129.4	0.245
D5A1T8	WP_006618695.1	toxin antitoxin system	4.45	8387.38	-0.526
D5A0X3	WP_014274422.1	toxin antitoxin system	9.03	4305.11	0.277
D4ZY60	WP_006619050.1	toxin antitoxin system	4.68	7131.15	-0.189
D4ZVC7	WP_006616204.1	toxin antitoxin system	5.01	15290.3	0.005
D4ZRN3	WP_006622250.1	toxin hicA family	4.04	5924.66	-0.039
D4ZYN0	WP_006616748.1	transcription termination factor	4.42	10734.89	-0.929

D4ZVT3	WP_006618212.1	transcriptional regulator	5.48	9328.51	-0.614
D4ZV55	WP_006617522.1	transcriptional regulator	9.22	7058.19	-0.603
D5A0R9	WP_006616065.1	transcriptional regulator	8.03	10419	-0.912
D4ZST6	WP_014275609.1	transcriptional regulator	4.77	121849.02	-0.291
D5A0X2	WP_006616235.1	transcriptional regulator	4.75	7222.21	-0.326
D4ZVD6	WP_014276130.1	transcriptional regulator	4.88	7341.39	-0.621
D5A0G2	WP_006620165.1	transcriptional regulator	7.16	45737.94	-0.535
D4ZX15	WP_006617491.1	transcriptional regulator	8.69	19741.82	-0.185
D4ZT86	WP_006615903.1	transcriptional regulator	4.07	8373.36	-0.215
D5A600	WP_014277125.1	transcriptional regulator	4.68	33518.24	-0.707
D4ZPS9	WP_006616028.1	transcriptional regulator	4.36	9471.58	-0.583
D4ZZT5	WP_014276559.1	transcriptional regulator	4.95	33137.82	-0.808
D4ZVT1	WP_006618214.1	transcriptional regulator	4.57	9147.54	0.014
D4ZYN4	WP_014276488.1	transcriptional regulator	4.92	32257.94	-0.721
D4ZP34	WP_006618619.1	transcriptional regulator	4.91	20315.9	-0.238
D4ZU92	WP_014275884.1	transcriptional regulator	5.61	60495.79	-0.812
D5A1U5	WP_014276781.1	transcriptional regulator	5.46	29657.81	-0.208
D4ZU85	WP_006618023.1	transcriptional regulator	8.45	8212.66	-0.306
D4ZVC8	WP_006616203.1	transcriptional regulator	4.53	7674.75	-0.146
D4ZWF8	WP_014273964.1	transcriptional regulator	8.49	15365.39	-0.839
D4ZZU9	WP_006617970.1	acetylase transferase	4.88	24783	-0.467
D5A3M9	WP_014274697.1	acetylglucosamine transferase	4.86	85918.76	-0.229
D5A3M8	WP_014274696.1	acetylglucosamine transferase	4.75	84603.57	-0.301
D5A268	WP_014274542.1	acyltransferase protein	6.34	33767.93	-0.239
D4ZRQ9	WP_014275433.1	arabinose transferase	9.29	72501.64	0.563
D5A6G4	WP_006616076.1	dna methyltransferase	4.92	26576.96	-0.136
D5A0M0	WP_006619577.1	dna methyltransferase	8.73	11505.19	-0.267
D4ZW45	WP_014276287.1	FkbM family methyltransferase	4.74	27205.95	-0.159
D5A2Q9	WP_014274660.1	glycerate kinase	5.11	40201.94	-0.488
D4ZNQ2	WP_006619527.1	glycerol acyltransferase	7.76	51290.05	-0.2
D4ZWN0	WP_006618846.1	glycogen debranching protein	4.41	7833.87	-0.356
D5A3C9	WP_006616048.1	glycosyl transferase	4.49	29117.68	-0.439
D5A2U3	WP_014276824.1	glycosyl transferase	5.31	41383.56	-0.039
D4ZVW4	WP_006620103.1	glycosyl transferase	6.2	60445.68	0.459
D4ZV89	WP_006618475.1	methyltransferases	5.48	33414.2	-0.199
D4ZQF5	WP_014277491.1	methyltransferases	6.47	90509.9	-0.38
D4ZPF5	WP_014275223.1	nucleotidyltransferase	4.75	32847.27	-0.168
D4ZW99	WP_006615910.1	phospho lactate guanyltltransferase	4.2	8087.14	-0.091
D4ZQC7	WP_006619132.1	prenyltransferase	9.25	38946.85	0.336
D5A4F7	WP_006617852.1	prenyltransferase	5.34	30345.46	-0.14
D5A5M8	WP_014274977.1	serine kinase	6.52	16433.66	0.141
D4ZTN5	WP_006620014.1	serine kinase	7.77	40270.5	-0.394
D5A500	WP_014277086.1	serine protein kinase	5.14	40637.49	-0.399
D5A4Z1	WP_014277080.1	serine protein kinase	5.79	49091.94	-0.359
D5A4Z0	WP_014277079.1	serine protein kinase	5.04	52855.79	-0.373
D5A316	WP_006619769.1	sugar transferase	8.17	39570.12	-0.308
D5A318	WP_006619767.1	sugar transferase	5.72	37556.64	-0.486
D4ZMX9	WP_006616561.1	sulfotransferase	4.75	19319.84	-0.231
D5A604	WP_014277128.1	xyloglucan fucosyltransferase	10.12	6530.79	-0.007
D5A1T4	WP_014276775.1	FtsX like permease	4.96	45054.18	0.3
D5A1D4	WP_006618107.1	ABC transporter	4.9	8482.86	0.392
D4ZU87	WP_006618025.1	ABC transporter	4.94	22449.57	-0.233
D4ZZX3	WP_006619180.1	ABC transporter	11.32	8836.62	0.705
D4ZZT8	WP_014276560.1	ABC transporter	4.64	65801.38	-0.22
D4ZXZ4	WP_014274117.1	ABC transporter	8.87	37228.82	-0.476
D4ZWA4	WP_014273928.1	ABC transporter	5.5	24625.81	-0.712
D4ZW79	WP_014273909.1	ABC transporter	5.51	26246.5	-0.806
D4ZPC9	WP_014275209.1	ABC transporter	4.73	49217.81	-0.103
D4ZRB4	WP_006619657.1	ABC transporter	9.36	27404.59	-0.257
D5A2J8	WP_006616649.1	carbohydrate porin	4.77	53915.17	0.017
D4ZPJ9	WP_014275233.1	sodium calcium exchange	3.59	160792.45	-0.634
D4ZQ91	WP_014277455.1	sodium calcium exchange	4.15	108796.59	-0.283
D5A209	WP_006619693.1	sodium hydrogen antiporter	4.45	10317.24	0.952
D5A208	WP_006619692.1	sodium hydrogen antiporter	5.98	9557.69	1.031
D4ZXB8	WP_014276370.1	sodium potassium transporter	3.7	155182.11	-0.088
D4ZVW9	WP_014276232.1	sodium potassium transporter	6.57	80245.95	0.501
D5A3S4	WP_014274728.1	sugar transporter	5.72	55861.64	-0.004
D4ZUD4	WP_014275911.1	thiamine transporter protein	9.23	41310.43	0.245
D4ZXU5	WP_014274088.1	transposase	10.37	8259.76	-0.473
D4ZY44	WP_006617395.1	transposase	6.08	13014.93	-0.353
D4ZNE3	WP_014277245.1	transposase	4.4	7683.88	0.169

D4ZXS2	WP 014274081.1	transposase	9.32	20363.42	-0.481
D4ZNC7	WP 014275154.1	transposase	4.74	10254.33	-0.884
D4ZNC9	WP 014275156.1	transposase	5.02	8925.07	-0.549
D4ZTZ9	WP 014275832.1	transposase	5.75	10793.57	-0.175
D5A0U0	WP 014274393.1	transposase	9.31	48628.8	-0.449
D4ZP87	WP 014275184.1	transposase	9.35	48586.82	-0.436
D4ZUI9	WP 014275940.1	transposase	9.75	38354.39	-0.353
D4ZYM4	WP 014276483.1	transposase	9.35	48577.81	-0.44
D4ZYZ5	WP 014274811.1	transposase	9.35	48572.73	-0.453
D5A1Q6	WP 014276758.1	transposase	7.94	8011.92	-0.638
D4ZY79	WP 006618650.1	addiction molecule toxin ReiE	8.58	14191.37	-0.392
D4ZUC4	WP 006616730.1	centromere protein	6.08	64601.01	-0.48
D5A0Q0	WP 014274369.1	circadian clock protein	5.11	17750.97	-0.444
D4ZRD0	WP 014277521.1	circadian oscillating protein	6.17	29564.47	-0.313
D5A0R1	WP 014274378.1	competence protein ComE	5.32	58511.38	-0.321
D4ZRL1	WP 006619855.1	competence protein comFB	9.38	17449.89	-0.794
D5A4L8	WP 006620033.1	competence protein ComFB	9.51	21933.03	-0.79
D5A3Y2	WP 014274770.1	CoxE protein	6.13	44993.75	-0.45
D4ZP18	WP 006618635.1	cytotoxic translational repressor	9.52	10117.86	0.045
D4ZYZ3	WP 006616328.1	ferredoxin domain protein	6.68	8557.86	0.246
D4ZVV9	WP 006620108.1	flxA like protein	4.88	21955.87	-0.615
D4ZZJ3	WP 006620329.1	glucose inhibited division protein	4.96	7401.58	0.593
D4ZNJ2	WP 006618959.1	glyoxalase domain protein	6.04	22470.06	0.081
D4ZZM3	WP 014274328.1	HEAT repeat domain protein/CpeF	4.53	36658.39	-0.268
D4ZNL6	WP 014277281.1	iron-sulfur cluster binding domain	8.23	65945.42	-0.39
D5A4S8	WP 014277051.1	LamG domain protein	5.2	262904.4	-0.484
D4ZX90	WP 014276356.1	low co2 inducible protein	5.7	27284.36	-0.256
D4ZUH0	WP 014275928.1	microcompartments protein	5.95	25903.64	0.087
D4ZN05	WP 014275091.1	neugrin	5.92	28294.18	-0.776
D5A0T6	WP 006616019.1	NTPase protein	10.4	16738.14	-0.348
D4ZVE1	WP 014276133.1	NTPase protein	5.18	18712.04	-0.747
D5A3J2	WP 014276952.1	NTPase protein	4.76	7492.62	-0.273
D4ZVW0	WP 006620107.1	ParM/StbA protein	6.05	40829.44	-0.193
D5A499	WP 014274842.1	patatin	6.71	8112.39	-0.633
D4ZNW8	WP 014277334.1	patatin	5.17	74934.67	-0.33
D4ZRW7	WP 006618597.1	peptidoglycan binding protein	5.82	45409.86	-0.203
D4ZZS6	WP 006619747.1	peroxiredoxin	6.46	15592.74	0.092
D5A3Q6	WP 014274717.1	polymerase	9.4	47553.68	0.777
D4ZXZ9	WP 006617784.1	prevent host death family protein	4.56	8476.79	-0.103
D4ZVX5	WP 006620093.1	ribosomal protein	8.71	10935.75	-0.15
D5A1C8	WP 014276694.1	RloB like protein	6.39	26707.1	-0.741
D4ZVX9	WP 006620089.1	secretine protein	4.4	19051.03	-0.482
D4ZRI5	WP 006618121.1	secretine protein	5.68	76281.88	-0.608
D4ZXZ5	WP 014274118.1	septum formation	5.43	42323.46	-0.667
D4ZXV8	WP 014274099.1	serine phosphatase	4.88	74682.87	-0.175
D4ZSW5	WP 014275629.1	spore germination protein	5.27	20311.08	-0.173
D5A5T7	WP 006617090.1	sxtJ protein	10.2	15560.5	0.312
D4ZNL2	WP 006617923.1	wd40 domain protein	4.68	56633.3	-0.269
D4ZRT7	WP 006618519.1	zinc finger protein	6.37	7476.63	0.172
D4ZQ73	WP 014277440.1	zinc ribbon domain protein	5.45	126678.28	-0.611

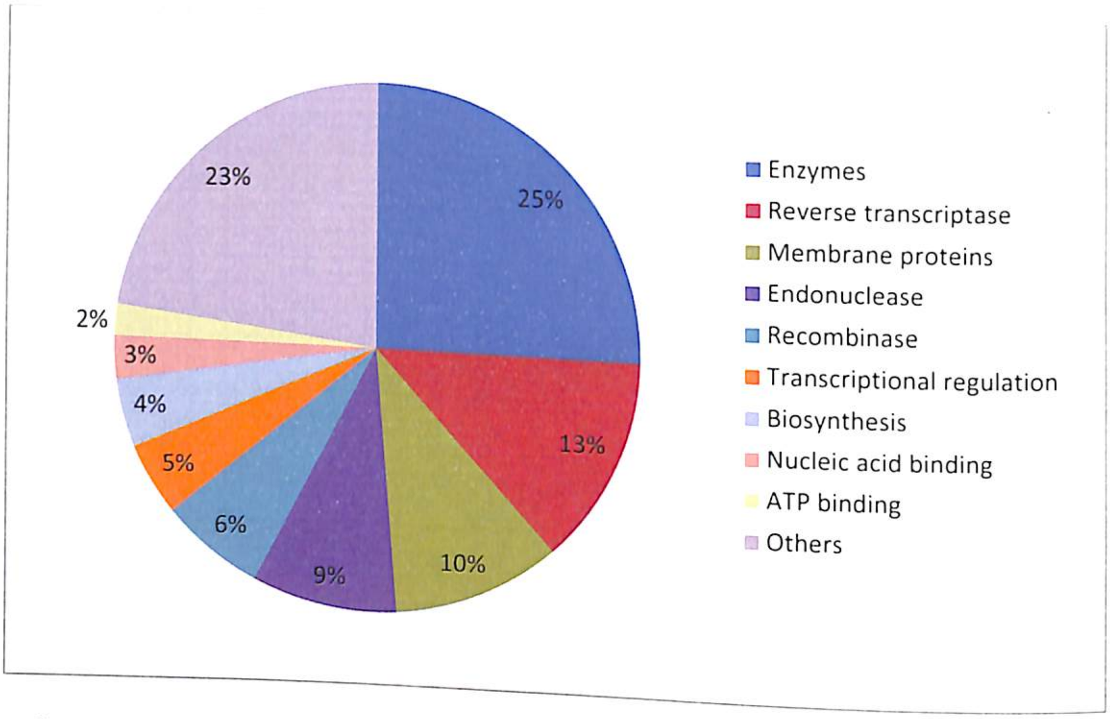


Figure 3.1 The annotated hypothetical proteins were divided into ten functional categories. Different enzymes except reverse transcriptase, endonucleases, recombinase are listed under enzyme category.

3.3.1.1 Enzymes

These are the enzymes that perform various catalytic activities. In our analysis, we found enzymes from different categories which constitute 25% of the total annotated proteins. These enzymes belonged to different classes of enzymes like aldolase, dehydrogenase like sterol desaturase, DNA helicase, peptidase, transferase like acetylase transferase, Oxidoreductases like methanol dehydrogenase, Hydrolases like serine hydrolase and hydrolyzing enzymes like chitinase. All these enzymes help in the normal functioning of the cell.

3.3.1.2 Reverse Transcriptase

Reverse transcriptases are enzymes that converts RNA to cDNA. They are used for genome replication in RNA containing virus, i.e. retrovirus. 13% of our annotated proteins were reverse transcriptases. This supports that cyanobacteria also act as a host for viruses.

3.3.1.3 Membrane proteins

Membrane proteins were also identified in this study. As a known alkalophile, *Arthrospira platensis* NIES-39 is still under research for resolving the mysteries behind its alkalophilic nature. Na^{+}/H^{+} antiporters were shown to play a significant role in alkali tolerance. Na^{+}/H^{+} antiporters were also detected in this study. Hence it is not surprising that the membrane proteins may play a vital role in alkali tolerance. In this study, we have found out that these annotated membrane proteins belong to some important membrane proteins like ABC transporters, ompA protein, FtsX protein, transmembrane transport, carbohydrate transport, antiporter activity and proteins involved in cell communication.

3.3.1.4 Endonuclease

Endonucleases are important enzymes in a cell. They are used in wide applications like DNA repair and various biotechnological processes with restriction endonucleases. We annotated endonucleases like HNH endonucleases, restriction endonucleases and Uma2 family endonucleases.

3.3.1.5 Recombinase

Recombinases are the enzymes which facilitates the recombination process. In the case of bacteria, the recombinase helps in the DNA repair mechanisms. We have also annotated several recombinases in our annotation process.

3.3.1.6 Transcriptional regulators

Regulation is a very important process in a cell's life. Every process in the cell is regulated precisely. Transcriptional regulation is one of the first types of regulation that a cell implements to ensure a smooth and error-free transcription.

3.3.1.7 Biosynthetic reactions

It is necessary to make new molecules in a cell as old molecules get degraded. Biosynthetic pathways play a key role in this regard. They generate a number of important molecules. Our study revealed some protein involved in biosynthetic pathways. These are biotin carboxylase, glutamate biogenesis protein, arginyl tRNA synthetase and ATP synthase.

3.3.1.8 Nucleic Acid binding proteins

Proteins binding to DNA and RNA come under this class. DNA binding proteins are the proteins which bind to DNA and carry out various functions. In our study, several functions of these DNA binding proteins have been detected like methylation, DNA polymerase. These proteins help key cellular processes like DNA replication. RNA polymerase is an RNA binding protein and helps in the transcription process.

3.3.1.9 ATP binding proteins

A cell cannot survive without energy. Many proteins and enzymes depend on the availability of ATP for their normal function. We have also found some ATP binding proteins during our annotation. These types of proteins have a domain(s) that specifically binds to ATP for energy. In this study, we found proteins like ATPase, protease and kinase that depends on ATP for energy.

3.3.1.10 Others

There are many other proteins which were annotated during this study but could not be categorized under the above-mentioned categories. Those proteins were either few in numbers, or only a single protein was annotated. However, many important proteins were identified. These proteins include proteins like calcium ion binding, metal ion binding, CRISPR related proteins, metal resistance, photosynthesis-related, signal transduction, circadian clock protein and stress tolerance proteins.

3.3.2 Physicochemical Characterization

The physiochemical properties like pI, molecular weight and Grand average of hydropathy (GRAVY) of all the hypothetical proteins were calculated by ProtParam server of Expasy and are listed in table 3.2. These properties aid in defining the function of a protein like with pI we can think of the probable environment in which an enzyme can work. In our study, we have found that many proteins have high pI values. After examining the results, we have found that high pI values belong to a particular category of proteins. For example, reverse transcriptase has a pI range from 9 to 12, while LtrA protein (A reverse transcriptase) also has a pI range from 11 to 12, it is interesting that both the enzymes are RNA binding proteins. Hnh endonucleases has a pI range from 10.38 to 11.66 while recombinase ranges from 7.67 to 10. In case of low pI values Uma2 endonucleases has a pI range from 4.5 to 5. GRAVY tells about the hydrophobic nature of the protein. In our study, we have identified many membrane proteins and transporters that have high GRAVY values which adds to our annotation process, since membrane proteins and transporters have a high percentage of hydrophobic amino acids as they are embedded in the lipid bilayer.

3.3.3 Pathway identification of the annotated proteins

All the annotated proteins were considered for the possible pathway assignment from the KEGG database. However, very few were currently found to be associated with any possible pathway. We found only 15 annotated proteins to be associated with a single or several pathways (Table 3.3). We found a range of different pathways associations with our annotated proteins. The highest association is found with Kinases which are involved in many pathways like calcium signalling, sphingolipid signalling, carbon metabolism, antibiotic biosynthesis and several others. Some other pathways found are lipopolysaccharide biosynthesis, Phenylalanine biosynthesis, ABC transporters family and cell cycle pathways.

Table 3.3 Pathways were identified for the annotated protein. 15 annotated proteins could be related to some pathway(s). KEGG database was used to find possible pathways associated with the annotated proteins.

UniProt ID	Annotated Function	Associated pathways
D4ZTT8	diacylglycerol kinase	Calcium signalling pathway Apelin signalling pathway Phospholipase D signalling pathway Sphingolipid signalling pathway
D4ZTG7	beta-glucosidase	Other glycan degradation
D4ZZ79	Lipid-a-disaccharide synthetase	Lipopolysaccharide biosynthesis
D4ZP07	oxopronilase	Glutathione metabolism
D4ZW68	aldolase	Dioxin degradation Xylene degradation Phenylalanine metabolism
D4ZRT8	oxidoreductase	Seleno compound metabolism
D5A3S4	sugar transporter	Biofilm formation
D4ZUI2	creatininase family protein	Arginine and proline metabolism
D4ZND5	carboxypeptidase	ABC transporters
D4ZPC9	ABC transporter	ABC transporters Quorum sensing
D5A121	ribonuclease HI	DNA replication
D4ZQS7	RNA binding protein hfq	Quorum sensing Biofilm formation RNA degradation
D4ZQ65	DNA adenine methylase	Mismatch repair
D5A0B4	aspartyl protease	Cell cycle
D5A2Q9	glycerate kinase	Carbon, Glyoxylate and dicarboxylate metabolism Glycerolipid metabolism Biosynthesis of antibiotics

3.3.4 Protein interaction network

Proteins usually work in co-operation. They interact with each other for various functions to work normally. The proteins that we have annotated in this study belong to various classes and perform different functions. Hence it is highly likely that these proteins interact with each other for their functioning. So, we generated a protein interaction network between these annotated proteins using STRINGS database. This interaction network consists of 526 annotated proteins, however here we have only shown those proteins which are interacting with other proteins. These proteins make 522 nodes connected with 2411 edges (interactions). The p-value for this network was $1.0e^{-16}$. P-value is the probability value or significance value for a statistical model. A low p-value indicates that there is a little chance that the results have derived from a chance.

This protein interaction map was divided into two main clusters. The cluster on the right side is a cluster of reverse transcriptase and LtrA proteins (a kind of reverse transcriptase). These enzymes have a high pI values and thus might be clubbed together. The one on the left contains other enzymes like restriction endonuclease, HNH endonucleases and other enzymes, that have low pI values. It also contains other predicted functional categories. Several pairwise interactions have been also seen in this interaction maps. Pairwise interactions are easy to analyse and hence we have investigated a pairwise interaction i.e. BAI93621.1 and BAI93619.1 from our protein interaction map. Our annotation linked BAI93621.1 to tellurite resistance while BAI93619.1 is linked to nucleotide pyrophosphatase. Tellurite resistance genes help in the efflux of tellurium ions from the cell. Nucleotide pyrophosphate might help in the hydrolysis of nucleotides like ATP which provides energy for the transport process.

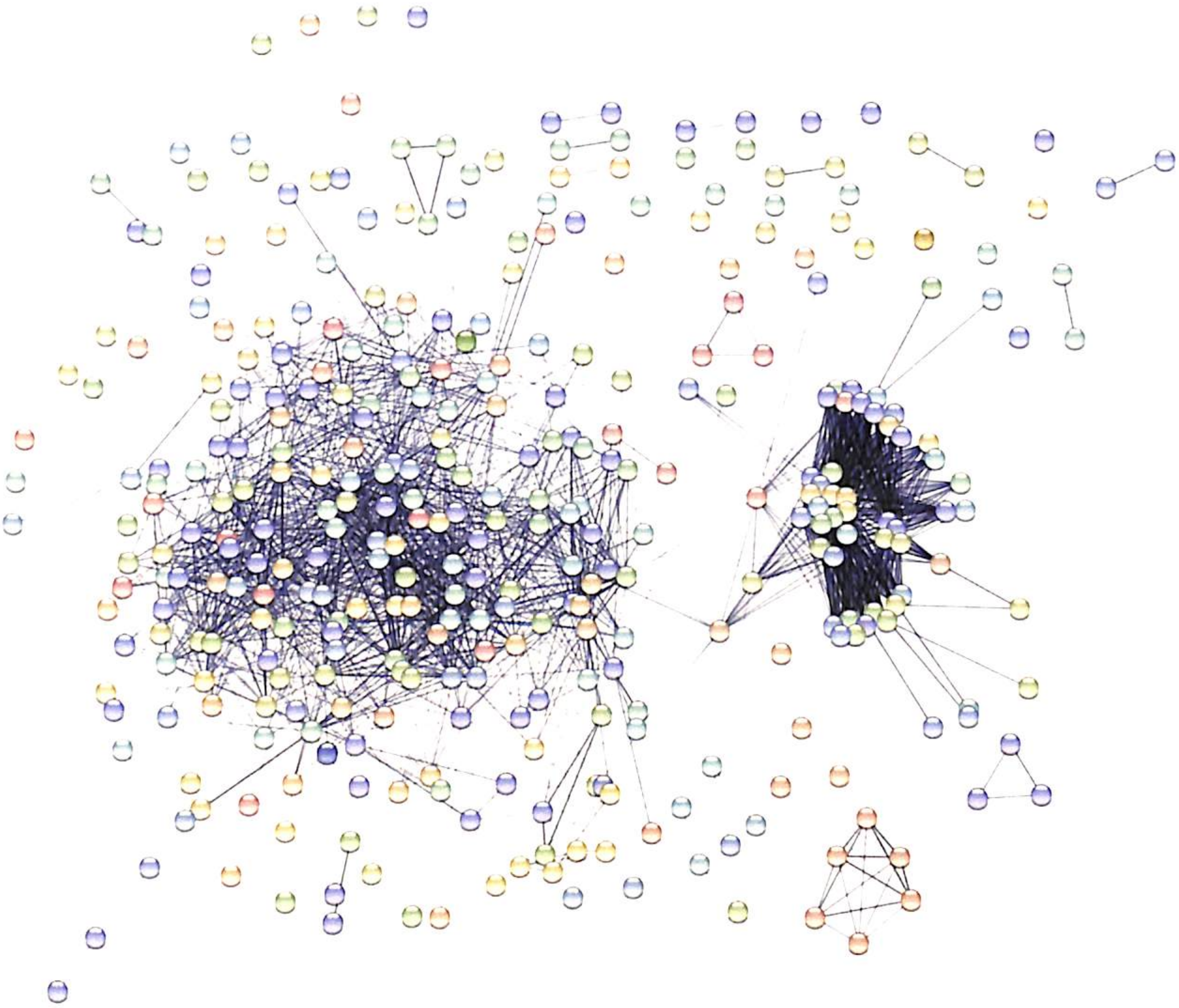


Figure 3.2 Protein interaction networks of the annotated proteins. The thickness of the line indicates the confidence of interaction. For these interactions, 522 proteins have been considered. These proteins make 522 nodes connected with 2411 edges (interactions). The p-value for this network is $1.0e^{-16}$.

3.4 Conclusions

Annotating a gene/protein sequence could lead us to a comprehensive understanding of the cellular working in terms of functional parameters. In the present study, we have annotated the currently un-annotated proteins of *Arthrospira platensis* NIES-39. Out of total 1364 un-annotated proteins, we were able to annotate 526 proteins. These 526 proteins belong to 10 different functional categories *viz.* enzymes, reverse transcriptase, membrane proteins, endonuclease, recombinase, transcriptional regulators, biosynthetic reactions, nucleic acid binding proteins, ATP binding proteins and other proteins. These categories contain some important proteins which we were able to annotate like ABC transporters, transcriptional regulators, restriction endonucleases, metal ion binding and many other functionally important enzymes. Out of these 526 annotated proteins, few proteins are found to be stress induced proteins like alpha crystalline family protein and nirD stress tolerance protein. While some proteins were also associated with the protein production machinery like many peptidases, chaperons, amino acids metabolism and a nitrate reductase associated protein. Annotated proteins were also assigned to several pathways like calcium signalling pathway, Apelin signalling pathway, biofilm formation, DNA replication, RNA degradation and cell cycle. Protein interaction network was also generated for the annotated proteins which showed high level of interaction between these proteins.

Chapter IV

Comparative analysis of Nitrate assimilatory enzymes among cyanobacteria

4.1 Introduction

All Nitrate Reductases (Prokaryotic and Eukaryotic) belong to the Molybdopterin-Binding (MopB) superfamily of proteins. MopB domain binds molybdopterin as a cofactor and has been reported in a variety of molybdenum and tungsten-containing enzymes, like formate dehydrogenase-H (Fdh-H) and -N (Fdh-N), some nitrate reductase (Nap, Nas, NarG), dimethylsulfoxide reductase (DMSOR), thiosulfate reductase, formylmethanofuran dehydrogenase, and arsenite oxidase (Maia and Moura 2015). Depending on the functions and organisms these proteins can exist in various forms like monomers, heterodimers, or heterotrimers. Cyanobacterial nitrate reductases are molybdoenzymes that catalyse the two-electron reduction of nitrate to nitrite. In cyanobacteria, NR contains the bis-molybdopterin guanine dinucleotide (bis-MGD) cofactor and a [3Fe-4S] cluster (Rubio et al. 1998, Rubio et al. 1999, Rubio et al. 2002).

Nitrite reductase belongs to the NIR_SIR_ferr superfamily. Sulfite and Nitrite reductases are key to biosynthetic assimilations of both sulfur and nitrogen and dissimilation of oxidised anions for energy transduction. Two copies of this repeat are found in Nitrite and Sulfite reductases and form a single structural domain. NiR converts nitrite to ammonium by a six-electron reduction mechanism (Knaff and Hirasawa 1991).

The main aim of this study is to find the putative role of nitrate reductase and nitrite reductase in the high protein content of the cyanobacterium *Arthrospira platensis* NIES-39. In this study, we are trying to decipher the sequence and structural features of these enzymes unique to *Arthrospira platensis* NIES-39 by comparing it with the other species of cyanobacterial class. In this comparison, we have considered the evolutionary approach as well as the sequence motif and structural domains across all cyanobacteria. In an evolutionary approach, we compared the 16s based species tree with that of gene/protein tree and looked that whether the gene/protein has evolved in a similar or in a different fashion to that of species evolution. We have also analysed the functionally important residues of these proteins in *Arthrospira platensis* NIES-39 to look for possible variations that could lead to any functional variation and hence contribute to higher protein content. Structural analyses were also performed to investigate any possible structural variations.

4.2 Materials and methods

4.2.1 Selection of cyanobacterial species

NCBI genome database (https://www.ncbi.nlm.nih.gov/genome) was used to list all the cyanobacteria which have been sequenced. The genomes which were present in the complete and the chromosome levels only were considered for further study. In the complete and the chromosome level, a total of 124 cyanobacterial species were present (June 2017). These 124 species were reconsidered to remove different strains of the same species. For the final selection of species, a species tree based on 16s rRNA gene sequences was constructed (Figure 4.1) for the selected 124 species. After implementing these changes, a total of 56 species had been selected for further analysis which belonged to 8 orders of cyanobacteria. An order-wise selection of initial as well as final species is given in table 4.1.

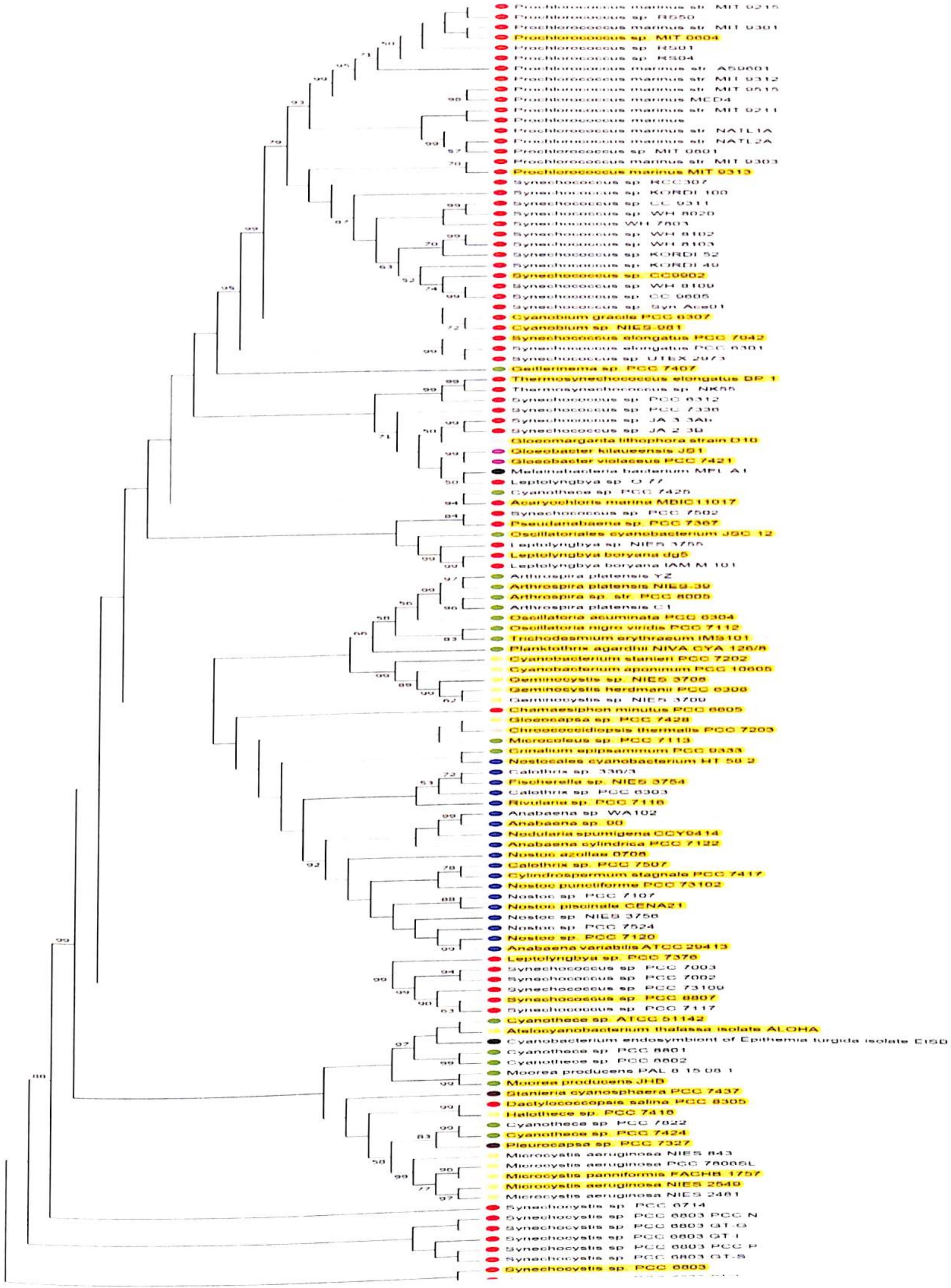

Prochlorococcus marinus str MIT 9215
Prochlorococcus sp RS50
Prochlorococcus marinus str MIT 9301
Prochlorococcus sp MIT 0604
Prochlorococcus sp RS01
Prochlorococcus sp RS04
Prochlorococcus marinus str AS9601
Prochlorococcus marinus str MIT 9312
Prochlorococcus marinus str MIT 9515
Prochlorococcus marinus MED4
Prochlorococcus marinus str MIT 9211
Prochlorococcus marinus
Prochlorococcus marinus str NATL1A
Prochlorococcus marinus str NATL2A
Prochlorococcus sp MIT 0801
Prochlorococcus marinus str MIT 9303
Prochlorococcus marinus MIT 9313
Synechococcus sp RCC307
Synechococcus sp KORDI 100
Synechococcus sp CC 9311
Synechococcus sp WH 8020
Synechococcus WH 7803
Synechococcus sp WH 8102
Synechococcus sp WH 8103
Synechococcus sp KORDI 52
Synechococcus sp KORDI 49
Synechococcus sp CC9902
Synechococcus sp WH 8109
Synechococcus sp CC 9605
Synechococcus sp Syn Ace01
Cyanobium gracile PCC 6307
Cyanobium sp NIES 981
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Synechococcus sp UTEX 2973
Geitlerinema sp PCC 7407
Thermosynechococcus elongatus BP 1
Thermosynechococcus sp NK55
Synechococcus sp PCC 6312
Synechococcus sp PCC 7336
Synechococcus sp JA 3 3Ab
Synechococcus sp JA 2 3B
Gloeomargarita lithophora strain D10
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Melainabacteria bacterium MEL A1
Leptolyngbya sp O 77
Cyanothece sp PCC 7425
Acaryochloris marina MBIC11017
Synechococcus sp PCC 7502
Pseudanabaena sp PCC 7367
Oscillatoriales cyanobacterium JSC 12
Leptolyngbya sp NIES 3755
Leptolyngbya boryana dg5
Leptolyngbya boryana IAM M 101
Arthrospira platensis YZ
Arthrospira platensis NIES 39
Arthrospira sp str PCC 8005
Arthrospira platensis C1
Oscillatoria acuminata PCC 6304
Oscillatoria nigro viridis PCC 7112
Trichodesmium erythraeum IMS101
Planktothrix agardhii NIVA CYA 126/8
Cyanobacterium stanieri PCC 7202
Cyanobacterium aponinum PCC 10605
Geminocystis sp NIES 3708
Geminocystis herdmanii PCC 6308
Geminocystis sp NIES 3709
Chamaesiphon minutus PCC 6605
Gloeocapsa sp PCC 7428
Chroococcidiopsis thermalis PCC 7203
Microcoleus sp PCC 7113
Crinalium epipsammum PCC 9333
Nostocales cyanobacterium HT 58 2
Calothrix sp 336/3
Fischerella sp NIES 3754
Calothrix sp PCC 6303
Rivularia sp PCC 7116
Anabaena sp WA102
Anabaena sp 90
Nodularia spumigena CCY9414
Anabaena cylindrica PCC 7122
Nostoc azollae 0708
Calothrix sp PCC 7507
Cylindrospermum stagnale PCC 7417
Nostoc punctiforme PCC 73102
Nostoc sp PCC 7107
Nostoc piscinale CENA21
Nostoc sp NIES 3756
Nostoc sp PCC 7524
Nostoc sp PCC 7120
Anabaena variabilis ATCC 29413
Leptolyngbya sp PCC 7376
Synechococcus sp PCC 7003
Synechococcus sp PCC 7002
Synechococcus sp PCC 73109
Synechococcus sp PCC 8807
Synechococcus sp PCC 7117
Cyanothece sp ATCC 51142
Atelocyanobacterium thalassa isolate ALOHA
Cyanobacterium endosymbiont of Epithemia turgida isolate EtSB
Cyanothece sp PCC 8801
Cyanothece sp PCC 8802
Moorea producens PAL 8 15 08 1
Moorea producens JHB
Stanieria cyanosphaera PCC 7437
Dactylococcopsis salina PCC 8305
Halothece sp PCC 7418
Cyanothece sp PCC 7822
Cyanothece sp PCC 7424
Pleurocapsa sp PCC 7327
Microcystis aeruginosa NIES 843
Microcystis aeruginosa PCC 7806SL
Microcystis panniformis FACHB 1757
Microcystis aeruginosa NIES 2549
Microcystis aeruginosa NIES 2481
Synechocystis sp PCC 6714
Synechocystis sp PCC 6803 PCC N
Synechocystis sp PCC 6803 GT-G
Synechocystis sp PCC 6803 GT-I
Synechocystis sp PCC 6803 PCC P
Synechocystis sp PCC 6803 GT-S
Synechocystis sp PCC 6803

Figure 4.1 16s rRNA gene sequences-based species tree for 124 cyanobacterial species. Highlighted species were selected for further analysis. Coloured circle represent the order of the species. Red colour represents Synechococcales; blue colour is Nostocales, green colour is Oscillatorials, yellow colour is Chrococcales, pink represents Gloebacterials, maroon represents Pleurocapsales, black and grey represent Chroococcidiopsidales and Gloeoemargaritales respectively.

4.2.2 Retrieval of nitrate reductase and nitrite reductase protein homologs

Nitrate reductase and nitrite reductase proteins from *Arthrospira platensis* NIES-39 were used as a query to retrieve the homologous protein in the selected 56 species from the National Center for Biotechnology Information (NCBI) RefSeq database. Blastn and Blastp (Basic Local Alignment Search Tool) (Altschul et al. 1990) were used against RefSeq protein database, and the organism was set to cyanobacteria (taxid 1117) for retrieving the homologous sequences of genes and the proteins from NCBI (with E-value cut off of ≤ 1 X 10^{-5}). In case of NR, 53 out of 56 homologs were retrieved. Sequences from *Prochlorococcus marinus* str. MIT 9313, *Nostoc azollae* 0708 and *Atelocyanobacterium thalassa* isolate ALOHA were not found. In the case of NiR 54 homologs were retrieved. The sequences from *Nostoc azollae* 0708 and *Atelocyanobacterium thalassa* isolate ALOHA were not found. The accession numbers of all the retrieved homologs are given in table 4.2.

Table 4.1 The number of species selected from each of the cyanobacterial order. Total 56 species were selected for current analysis.

S.No.	Order	Initial Number of species	Final Number of species
1.	Synechococcales	64	15
2.	Oscillatorials	20	13
3.	Nostocales	19	13
4.	Chrococcales	13	09
5.	Gloebacterials	02	02
6.	Pleurocapsales	02	02
7.	Chroococcidiopsidales	01	01
8.	Gloeoemargaritales	01	01
9.	Unidentified	02	00
	TOTAL	**124**	**56**

Table 4.2 Genome assembly number and the protein accession number of the selected NR and NiR enzymes.

S.No.	Organism Name	Order	Assembly	Protein Accession	
				NR	NiR
1	*Acaryochloris marina* MBIC11017	Synechococcales	GCA_000018105.1	WP_012163416.1	WP_041659830.1
2	*Chamaesiphon minutus* PCC 6605	Synechococcales	GCA_000317145.1	WP_015160099.1	WP 015161463.1
3	*Cyanobium gracile* PCC 6307	Synechococcales	GCA_000316515.1	WP_015109988.1	WP_043325795.1
4	*Cyanobium* sp. NIES-981	Synechococcales	GCA_900088535.1	WP_087068507.1	WP_087068510.1
5	*Dactylococcopsis salina* PCC 8305	Synechococcales	GCA_000317615.1	WP_015230744.1	WP_015230746.1
6	*Leptolyngbya boryana* dg5	Synechococcales	GCA_002142495.1	WP_017288929.1	WP_017288935.1
7	*Leptolyngbya* sp. PCC 7376	Synechococcales	GCA_000316605.1	WP_015132957.1	WP_015134937.1
8	*Prochlorococcus marinus* str. MIT 9313	Synechococcales	GCA_000011485.1	NA	WP_011131603.1
9	*Prochlorococcus* sp. MIT 0604	Synechococcales	GCA_000757845.1	WP_042851326.1	WP_042850618.1
10	*Pseudanabaena* sp. PCC 7367	Synechococcales	GCA_000317065.1	WP_041699619.1	WP_015165563.1
11	*Synechococcus elongatus* PCC 7942	Synechococcales	GCA_000012525.1	WP_011377931.1	WP_011242624.1
12	*Synechococcus* sp. CC9902	Synechococcales	GCA_000012505.1	WP_011361006.1	WP_011361018.1
13	*Synechococcus* sp. PCC 8807	Synechococcales	GCA_001693295.1	WP_065716331.1	WP_065716665.1
14	*Synechocystis* sp. PCC 6803	Synechococcales	GCA_000340785.1	WP_010872118.1	WP_010873675.1
15	*Thermosynechococcus elongatus* BP-1	Synechococcales	GCA_000011345.1	NP_682145.1	NP_682139.1
16	*Arthrospira platensis* NIES-39	Oscillatorials	GCA_000210375.1	WP_014274817.1	WP_014275660.1
17	*Arthrospira* sp. PCC 8005	Oscillatorials	GCA_000973065.1	WP_008049497.1	CDM94270.1
18	*Crinalium epipsammum* PCC 9333	Oscillatorials	GCA_000317495.1	WP_041226795.1	WP_015204592.1
19	*Cyanothece* sp. ATCC 51142	Oscillatorials	GCA_000017845.1	WP_009544043.1	WP_012361573.1

20	*Cyanothece* sp. PCC 7424	Oscillatorials	GCA_000021825.1	WP_015955450.1	WP_012599063.1
21	*Geitlerinema* sp. PCC 7407	Oscillatorials	GCA_000317045.1	WP_015170822.1	WP_015170815.1
22	*Microcoleus* sp. PCC 7113	Oscillatorials	GCA_000317515.1	WP_015180228.1	WP_015180235.1
23	*Moorea producens* JHB	Oscillatorials	GCA_001854205.1	WP_071103809.1	WP_071103805.1
24	*Oscillatoria acuminata* PCC 6304	Oscillatorials	GCA_000317105.1	WP_015152060.1	WP_015152056.1
25	*Oscillatoria nigro-viridis* PCC 7112	Oscillatorials	GCA_000317475.1	WP_015179103.1	WP_041623582.1
26	*Oscillatoriales cyanobacterium* JSC-12	Oscillatorials	GCA_000309945.1	WP_009554631.1	WP_009554639.1
27	*Planktothrix agardhii* NIVA-CYA 126/8	Oscillatorials	GCA_000710505.1	WP_042154875.1	WP_042154884.1
28	*Trichodesmium erythraeum* IMS101	Oscillatorials	GCA_000014265.1	WP_011610825.1	WP_011610823.1
29	*Anabaena cylindrica* PCC 7122	Nostocales	GCA_000317695.1	WP_096713173.1	WP_015213651.1
30	*Anabaena* sp. 90	Nostocales	GCA_000312705.1	WP_015081557.1	WP_015081559.1
31	*Anabaena variabilis* ATCC 29413	Nostocales	GCA_000204075.1	WP_011321213.1	WP_011321208.1
32	*Calothrix* sp. PCC 7507	Nostocales	GCA_000316575.1	WP_042341266.1	WP_015128366.1
33	*Cylindrospermum stagnale* PCC 7417	Nostocales	GCA_000317535.1	WP_041233704.1	WP_015209003.1
34	*Fischerella* sp. NIES-3754	Nostocales	GCA_001548455.1	WP_062247080.1	WP_062247088.1
35	*Nodularia spumigena* CCY9414	Nostocales	GCA_000340565.3	WP_006196196.1	WP_006196192.1
36	*Nostoc azollae* 0708	Nostocales	GCA_000196515.1	NA	NA
37	*Nostoc piscinale* CENA21	Nostocales	GCA_001298445.1	WP_062289963.1	WP_062289979.1
38	*Nostoc punctiforme* PCC 73102	Nostocales	GCA_000020025.1	WP_012408233.1	WP_012408235.1
39	*Nostoc* sp. PCC 7120	Nostocales	GCA_000009705.1	WP_010994788.1	WP_010994783.1
40	*Nostocales cyanobacterium* HT-58-2	Nostocales	GCA_002163975.1	WP_087537922.1	WP_087537929.1

41	*Rivularia* sp. PCC 7116	Nostocales	GCA_000316665.1	WP_015116720.1	WP_044290732.1
42	*Atelocyanobacterium thalassa* isolate ALOHA	Chroococcales	GCA_000025125.1	NA	NA
43	*Cyanobacterium aponinum* PCC 10605	Chroococcales	GCA_000317675.1	WP_041922971.1	WP_015218171.1
44	*Cyanobacterium stanieri* PCC 7202	Chroococcales	GCA_000317655.1	WP_015222182.1	WP_015222811.1
45	*Geminocystis herdmanii* PCC 6308	Chroococcales	GCA_000332235.1	WP_017294885.1	WP_017292467.1
46	*Geminocystis* sp. NIES-3708	Chroococcales	GCA_001548095.1	WP_066344784.1	WP_066347786.1
47	*Gloeocapsa* sp. PCC 7428	Chroococcales	GCA_000317555.1	WP_015188386.1	WP_015188381.1
48	*Halothece* sp. PCC 7418	Chroococcales	GCA_000317635.1	WP_015226432.1	WP_015224664.1
49	*Microcystis aeruginosa* NIES-2549	Chroococcales	GCA_000981785.1	WP_046662561.1	WP_046662719.1
50	*Microcystis panniformis* FACHB-1757	Chroococcales	GCA_001264245.1	AKV66046.1	AKV69016.1
51	*Pleurocapsa* sp. PCC 7327	Pleurocapsales	GCA_000317025.1	WP_015145301.1	WP_015142945.1
52	*Stanieria cyanosphaera* PCC 7437	Pleurocapsales	GCA_000317575.1	WP_015192771.1	WP_015192765.1
53	*Gloeobacter kilaueensis* JS1	Gloeobacterales	GCA_000484535.1	WP_023175789.1	WP_023171581.1
54	*Gloeobacter violaceus* PCC 7421	Gloeobacterales	GCA_000011385.1	NP_924517.1	NP_924503.1
55	*Chroococcidiopsis thermalis* PCC 7203	Chroococcidiopsidales	GCA_000317125.1	WP_015155574.1	WP_015157292.1
56	*Gloeomargarita lithophora* Alchichica-D10	Gloeoemargaritales	GCA_001870225.1	WP_071454570.1	WP_071454571.1

4.3 Results and Discussions

4.3.1Nitrate Reductase (NR)

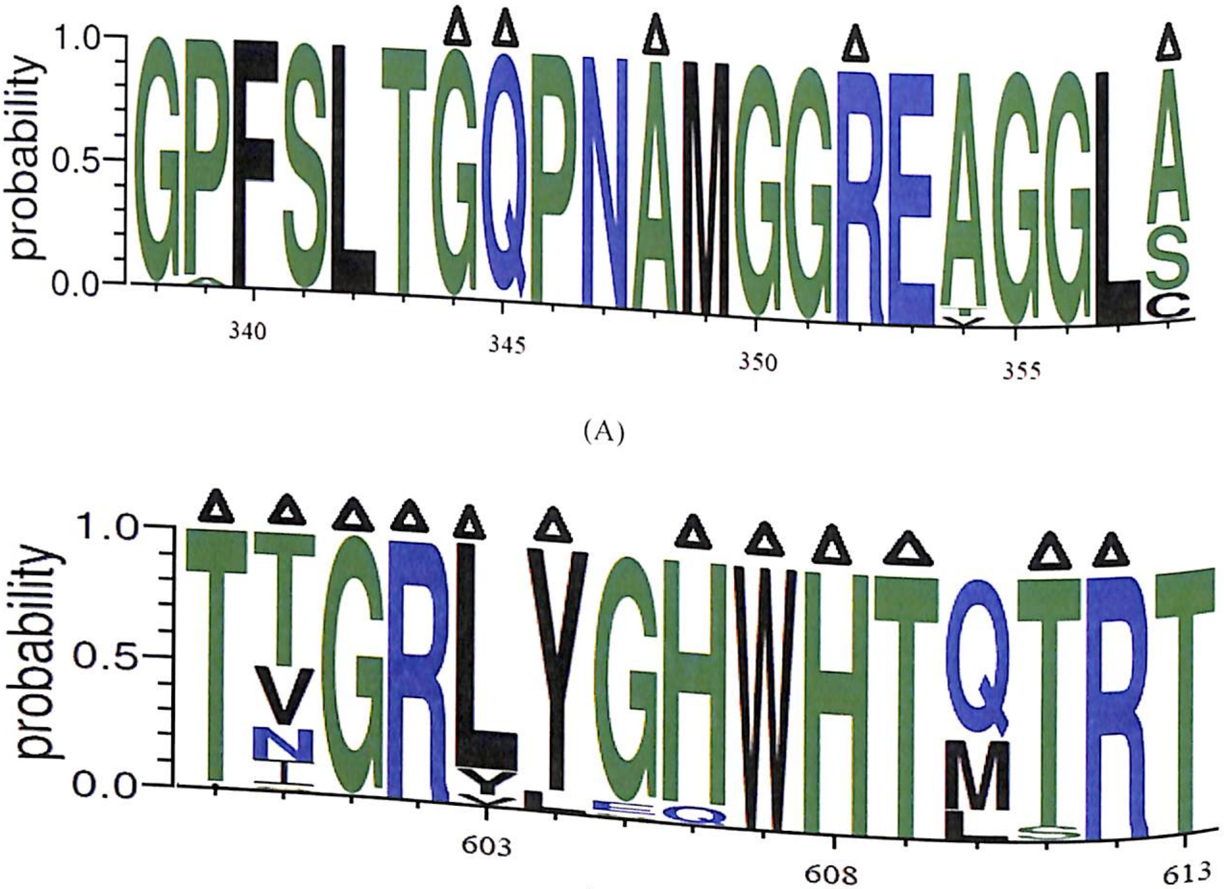
4.3.1.1 Sequence and structural analysis

All the cyanobacterial NR proteins belong to the Molybdopterin binding superfamily. The monomeric cyanobacterial NR protein comprises of two functional domains (Table 4.3) with an average 735 amino acid residues. The first domain is MopB Nitrate R NapA like (cd02754) and the second one is MopB CT Nitrate R NapA-like (cd02791) domains. MopB Nitrate R NapA-like domain contains an Iron-sulphur cluster binding region and a binding site for Molybdopterin cofactor. The much smaller MopB CT Nitrate R NapA-like domain is also involved in Molybdopterin cofactor binding. Molybdenum is coordinated by six ligands, out of six, four are provided by the two dithiolene sulfur atoms from two molybdopterin guanine dinucleotide (MGD) molecules, the fifth ligand is the Sulphur atom of Cys140 and the sixth is a non-proteineous sulphur. This Molybdenum cofactor helps in the transfer of electrons and formation of the intermediate during the enzymatic reactions. We identified the conserved signature pattern of these domains in cyanobacteria. MopB Nitrate R NapA-like domain is highly conserved, and a 21 amino acids long pattern is present in all the cyanobacteria which also have functionally important residues like G344 and Q345 which are involved in MGD2 binding while A348, R352 and A358 are involved in guiding the nitrate towards the active site (Figure 4.2A). MopB CT Nitrate R NapA-like domain is also conserved, although less than the first one. Its signature pattern consists of 15 residues with functionally importance such as T600, R602, W607, H608, T609, T611 and R612 which are involved in MGD1 binding while T599, G601, L603, Y604 and H606 are involved in MGD2 binding (Figure 4.2B). Among the 56 cyanobacterial species, *Prochlorococcus marinus* str. MIT 9313, *Nostoc azollae* 0708 and *Atelocyanobacterium thalassa* isolate ALOHA do not contain the NR gene or protein. The lack of NR in one out of these three species, i.e. *Prochlorococcus marinus* str. MIT 9313 is well documented (Dufresne et al. 2003, Rocap et al. 2003) as this strain may rely on reduced nitrogen compounds, such as NH_4^+ and amino acids for growth.

Table 4.3 Domains boundary of NR protein in each of the cyanobacterial species shows conserved nature of NR protein.

Domains	MopB_Nitrate-R-NapA-like (cd02754)			MopB_CT_Nitrate-R-NapA-like (cd02791)		
Species	From	To	Length	From	To	Length
Acaryochloris marina MBIC11017	6	589	584	597	715	119
Chamaesiphon minutus PCC 6605	6	584	579	591	711	121
Cyanobium gracile PCC 6307	16	590	575	598	717	120
Cyanobium sp. NIES-981	10	590	581	596	717	122
Dactylococcopsis salina PCC 8305	4	580	577	590	714	125
Leptolyngbya boryana dg5	6	589	584	595	716	122
Leptolyngbya sp. PCC 7376	7	599	593	606	726	121
Prochlorococcus sp. MIT 0604	5	568	564	582	702	121
Pseudanabaena sp. PCC 7367	5	569	565	583	702	120
Synechococcus elongatus PCC 7942	20	595	576	602	723	122
Synechococcus sp. CC9902	9	590	582	596	717	122
Synechococcus sp. PCC 8807	7	597	591	604	724	121
Synechocystis sp. PCC 6803	11	587	577	595	714	120
Thermosynechococcus elongatus BP-1	35	595	561	611	729	119
Arthrospira platensis NIES-39	6	580	575	593	710	118
Arthrospira sp. PCC 8005	6	580	575	593	710	118
Crinalium epipsammum PCC 9333	6	597	592	605	723	119
Cyanothece sp. ATCC 51142	6	594	589	600	718	119
Cyanothece sp. PCC 7424	6	605	600	611	732	122
Geitlerinema sp. PCC 7407	6	584	579	592	711	120
Microcoleus sp. PCC 7113	6	596	591	604	723	120
Moorea producens JHB	6	598	592	604	725	122
Oscillatoria acuminata PCC 6304	7	586	580	594	712	119
Oscillatoria nigro-viridis PCC 7112	5	574	570	582	701	120
Oscillatoriales cyanobacterium JSC-12	7	584	578	598	716	119
Planktothrix agardhii NIVA-CYA 126/8	5	585	581	592	709	118
Trichodesmium erythraeum IMS101	5	576	572	589	709	121
Anabaena cylindrica PCC 7122	6	579	574	593	712	120
Anabaena sp. 90	6	577	572	589	710	122
Anabaena variabilis ATCC 29413	6	578	573	592	711	120
Calothrix sp. PCC 7507	6	602	597	609	726	118
Cylindrospermum stagnale PCC 7417	6	608	603	615	735	121
Fischerella sp. NIES-3754	6	608	603	615	735	121
Nodularia spumigena CCY9414	6	588	583	601	721	121
Nostoc piscinale CENA21	6	586	581	600	719	120
Nostoc punctiforme PCC 73102	6	611	606	624	744	121
Nostoc sp. PCC 7120	6	597	592	611	730	120
Nostocales cyanobacterium HT-58-2	6	591	586	599	718	120
Rivularia sp. PCC 7116	6	607	602	614	734	121
Cyanobacterium aponinum PCC 10605	6	587	582	594	714	121
Cyanobacterium stanieri PCC 7202	6	577	572	584	704	121
Geminocystis herdmanii PCC 6308	6	585	580	592	711	120
Geminocystis sp. NIES-3708	6	585	580	592	712	121
Gloeocapsa sp. PCC 7428	6	592	587	599	716	118

Halothece sp. PCC 7418	4	580	577	592	712	121
Microcystis aeruginosa NIES-2549	6	596	591	602	723	122
Microcystis panniformis FACHB-1757	6	596	591	602	722	121
Pleurocapsa sp. PCC 7327	6	605	600	611	732	122
Stanieria cyanosphaera PCC 7437	6	608	603	615	735	121
Gloeobacter kilaueensis JS1	8	571	564	578	695	118
Gloeobacter violaceus PCC 7421	8	564	557	571	688	118
Chroococcidiopsis thermalis PCC 7203	6	589	584	597	715	119
Gloeomargarita lithophora Alchichica-D10	3	572	570	579	698	120



(A)

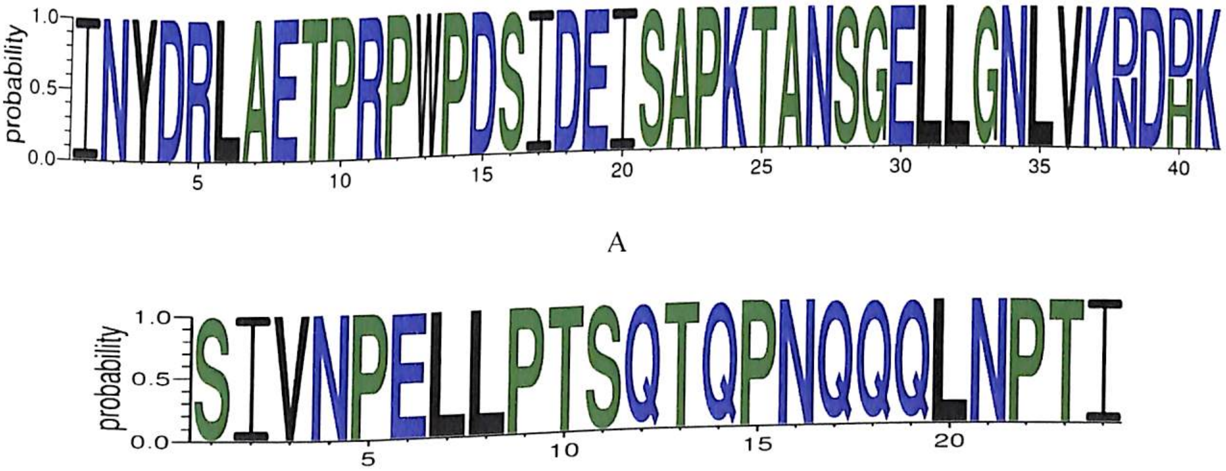
(B)

Figure 4.2 Conserved region of the NR domain in cyanobacteria (A) Signature pattern of MopB Nitrate R NapA like domain. The residues having a triangle on top are functionally important residues. G344 and Q345 are involved in MGD2 binding while A348, R352 and A358 are involved in the guiding of nitrate towards the active site (B) Conserved region of the MopB CT Nitrate R NapA-like domain. Functionally important residues T600, R602, W607, H608, T609, T611 and R612 are involved in MGD1 binding while T599, G601, L603, Y604 and H606 are involved in MGD2 binding.

Multiple sequence alignment (MSA) showed that the sequences are highly conserved with 262 residues showing more than 90% conservation in the total length of the protein. To identify any new potential motifs, we have used MEME program. MEME results identified ten major motifs in all the cyanobacterial sequences (Table 4.4). Interestingly, an extra motif of 24 amino acids (SIVNPELLPTSQTQPNQQQLNPTI) (E value = 6.9e-005) was identified in the genus *Arthropsira* at the C terminal and spans from residue 711 to 734. The sequence conservation for this motif is shown in terms of sequence logo in figure 4.3A. Another motif was identified in *Microcystis aeruginosa* NIES 2549 and *Microcystis panniformis* FACHB 1757 (WPDSIDEISAPKTANSGELLGNLVK[DN]D[HD]K) (E value =2.3e-006). This motif is 29 amino acids long and present from residue 536 to 564. The sequence conservation for this motif is shown in terms of sequence logo in figure 4.3B. These motifs are important regarding that these could be used as a potential marker to identify these Genera.

Table 4.4 MEME output showing the ten major motifs identified in the cyanobacterial species. Sites represent the number of sequences in which a particular motif was identified.

Motif	E-Value	Sites	Width
1	7.6e-3297	52	64
2	1.5e-5012	52	113
3	1.7e-2965	52	66
4	1.1e-3377	52	80
5	3.9e-2754	51	80
6	2.2e-2360	52	81
7	3.2e-2342	52	57
8	2.9e-2075	51	57
9	5.6e-1294	52	29
10	9.6e-873	50	29



A

B

Figure 4.3 Sequence conservations representing motifs in (A) Genus *Microcystis* of 41 amino acids and (B) Genus *Arthrospira* of 24 amino acids.

Arthrospira platensis NIES-39 protein sequence was used as a query to search Protein Data Bank. Bacterial periplasmic NR (Nap) from *Desulfovibrio desulfuricans* ATCC 27774 (PDB ID - 2JIO) (Najmudin et al. 2008) was the best hit with an identity of 35% and query coverage of 95%. Pairwise sequence alignment was performed between the bacterial and the cyanobacterial sequences and using the functionally important residues of bacteria the corresponding functionally important residues were identified in *Arthrospira platensis* NIES-39. A total of 70 functionally important residues were found conserved (Table 4.5). These 70 residues can be divided into five categories viz. (1) MGD1 binding – 26 residues (2) MGD2 binding – 25 residues (3) To guide nitrate into the active site – 9 residues (4) Iron Sulphur cluster binding – 9 residues (5) Molybdenum binding – 1 residue.

Table 4.5 Sequence variations (functionally important residues of NR protein) among cyanobacterial species. Functionally important residues were identified by comparing the sequences of *Desulfovibrio desulfuricans* ATCC 27774 and *Arthrospira platensis* NIES-39.

Desulfovibrio desulfuricans ATCC 27774	*Arthrospira platensis* NIES-39	*Synechocystis* sp. PCC 6803	Variations in Cyanobacteria	Type of substitution
		MGD 1 binding		
R 14	P10	P	P(52) S(1)	Differ
Q 111	Q119	Q		
N 136	N144	N		
Q 312	Q311	Q		
E 416	A414	A		
T 417	T415	T		
N 418	N416	N		
T 422	S420	S		
I 443	Q440	Q	Q(49) S(3) N(1)	Similar
E 444	D441	D	D(45) E(8)	Similar
A 445	A442	A	A(49) C(3) S(1)	Differ
F 446	Y443	Y		
P 461	A458	A	A(50) C(1)G(1)T(1)	Differ
A 462	A459	A	A(43) T(6) L(2) S(1) R(1)	Differ
F 463	Q460	Q		
S 615	T600	N	T(28) V(12) N(7)I(5) S(1)	Differ
R 617	R602	R		
W 622	W607	W		
H 623	H608	H		
T 624	T609	T		
T 626	T611	T	T(50) S(3)	Similar
M 627	R612	R		
F 689	M674	M		
N 697	N687	N		
Y 713	L703	L	L(52) V(1)	Similar
K 714	K704	K		

MGD 2 binding				
K 49	K58	K		
I 173	I181	I	I(44)V(9)	Similar
G 174	G182	G		
S 175	T183	T	T(46) S(6) A(1)	Similar
N 176	N184	N		
E 179	E187	D	E(46)D(6) A(1)	Similar
A 180	C188	C		
D 204	D212	D		
P 205	P213	P		
R 206	R214	R		
P 222	P230	P	P(51) S(1)L(1)	Differ
G 223	G231	G		
D 225	D233	D		
C 307	S306	S		
M 308	M307	M		
G 309	G308	G		
R 313	S312	S	S(51) R(2)	Differ
G 345	G344	G		
Q 346	Q345	Q		
T 614	T599	T	T(52) I(1)	Differ
M 616	G601	G		
V 618	L603	L	L(46) Y(5)V(2)	Differ
I 619	Y604	Y	Y(49) L(4)	Differ
H 621	H606	H	H(49) Q(4)	Differ
I 715	A705	A	A(52) S(1)	Differ
To guide nitrate into the active site				
R 138	R146	R		
R 354	R352	R		
S 360	A 358	S	A(33) S(14) C(6)	Differ
H 396	N 394	S	T(31) D(12) S(5) N(3) A(2)	Differ
Y 533	C 529	C		
R 709	K 699	K	L(26) K(17) G(4) R(3)H(1) S(1) Q(1)	Differ
P137	S145	S		
M141	M149	M		
A349	A348	A		
Iron Sulphur cluster binding				
C 13	C9	C		
Y 15	Y11	Y	Y(52) F(1)	Similar
C 16	C12	C		
G 19	G15	G		
C 20	C16	C		
C 47	C56	C		
G 50	G59	G		
P 182	P190	P		
V 183	I191	I	I(48) V(5)	Similar
Molybdenum binding				
C 140	C148	C		

All the above residues were analysed by looking into the MSA and searching for any type of variation within all the selected cyanobacterial species. We found that residues at only six positions were showing variations where the residue type was different in various species. Two residues belonged to the MGD1 binding category (459 and 600), one to MGD2 binding (603) while the remaining three belonged to the guiding nitrate towards active site (358, 394 and 699) (Numbered according to *Arthrospira platensis* NIES-39). The sequence conservations at these identified six positions are given in figure 4.4 in the form of sequence logos.

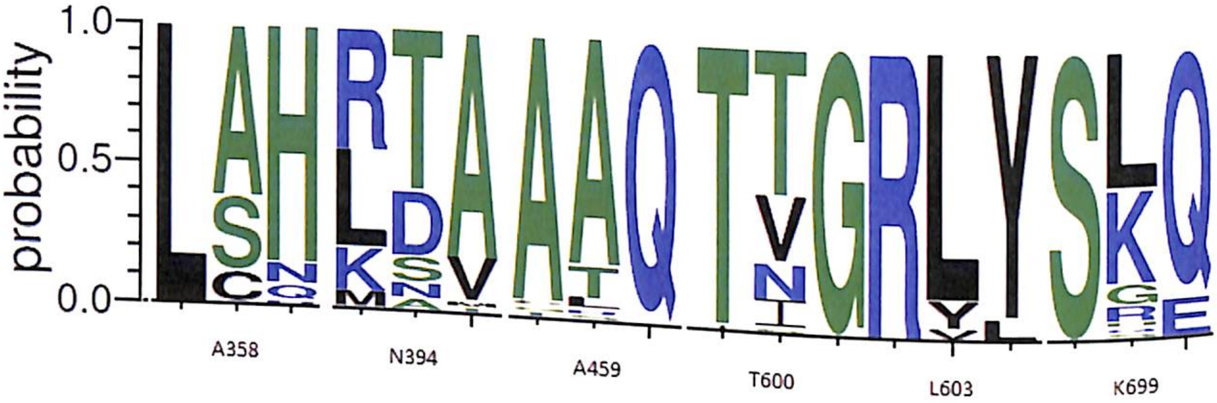


Figure 4.4 Sequence conservations at the functionally important residues in which variations were detected among cyanobacteria.

4.3.1.2 Phylogenetic analysis

4.3.1.2.1 Species tree

We generated 16s rRNA gene sequences-based species tree for all the selected 56 species for which the gene and protein were available (Figure 4.5). Our species tree was divided into a total of 12 clades. These 12 clades were distributed into 5 pure and 7 mixed clades. Pure clades had species from the same order while mixed clades had species from several orders. Pure clades included clade 1 and clade 7 which had species from the order Synechoccocales, clade 8 had Chrococcales, clade 9 had Oscillatorials and clade 12 had Nostocales. Rest of the 7 clades shared species from several orders.

This type of arrangement of species in the clades showed a mix-up for a few classical taxonomical orders, i.e. the classical orders were not included in the same clade. For instance, clade 2 has 5 species from 4 orders viz. Oscillatorials, Synechococcales, Gloeomargaritales and Gloeobacterials. This mix-up is quite common as all the current species are assigned to a taxonomical order based on morphological, physiological or biochemical data. It seems that this type of data is not sufficient to decipher the true evolutionary history of cyanobacteria.

Other studies have also shown mixed clades for cyanobacterial order (Gupta 2009, Singh et al. 2015). Our species tree is coherent with previously made 16s rRNA species tree using similar cyanobacterial genus (Seo and Yokota 2003, Shih et al. 2013). This species tree suggests that the order Nostocales is the most conserved order as the species of this order are confined to a single clade. Species within Oscillatorial is well conserved with a few exceptions. Chrococcales and Synechococcales are diverse orders since these orders comprise of species located within different clades in the tree with high bootstrap support values. This suggests that these orders were constantly interacting at a gene level. This species tree also indicates a general pattern of evolution of the cyanobacterial orders, i.e. constant gene duplication and speciation events took place from the first common ancestor of Cyanobacteria. We have considered our species tree as a reference tree and compared our gene/protein trees to this reference to get a picture of the protein evolution.

4.3.1.2.2 Gene tree

Both Maximum likelihood (ML) and Neighbour-joining (NJ) methods were used to generate the phylogenetic tree for the gene of NR (*narB*) using the 53 species (Figure 4.6). The topology of both the trees was similar, so only NJ tree has been discussed here. Our gene tree has been divided into 15 clades. Out of these 15 clades, 8 contain species from the same order, i.e. clade 1 and 4 include Synechococales, clade 3 and 13 include Chroccocales, clades 6, 10 and 15 include Oscillatorials and clade 9 include Nostocales. The remaining 7 clades contains species from different orders. For example, clade 11 has 4 species from 4 orders which are Synechococales, Nostocales, Oscillatorials and Pleurocapsales. Presence of these species from different orders in a single clade points towards a common evolution of this gene due to a similar kind of environmental pressure. A horizontal gene transfer event likely occurred between these species. To get a clear picture of the gene evolution, we compared our species tree with the gene tree. Clade wise comparison of species and gene tree revealed that while most of the species retained their clades with other co-species in gene tree as was in species tree, some species moved to an entirely different clade with different co-species. However, in most cases the bootstrap values of these branches are not significant. Only in the case of *Rivularia* sp. PCC, we observed that the evolutionary pattern of the gene is different from that of species. *Arthrospira platensis* NIES-39 was present in clade 10 with three other species of the order Oscillatorials which are *Arthrospira* sp. str. PCC 8005, *Trichodesmium erythraeum* and *Cyanothece* sp. ATCC 51142 indicating the closeness of NR gene of

Arthrospira platensis NIES-39 to NR gene of other Oscillatorials. Other members of the order Oscillatorials were present in different clades of the tree.

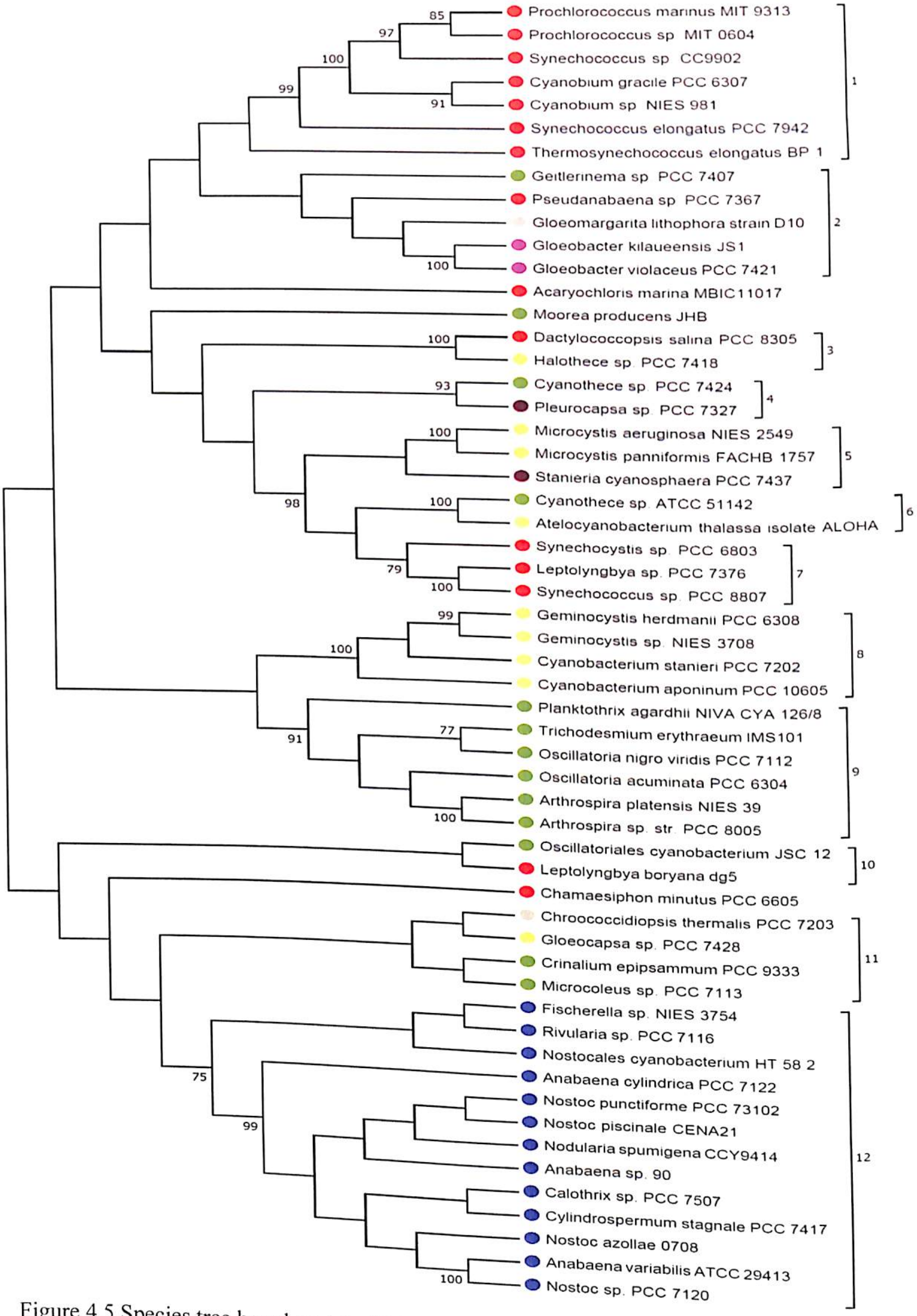

85
97
100
99
91
Prochlorococcus marinus MIT 9313
Prochlorococcus sp MIT 0604
Synechococcus sp CC9902
Cyanobium gracile PCC 6307
Cyanobium sp NIES 981
Synechococcus elongatus PCC 7942
Thermosynechococcus elongatus BP 1
1
Geitlerinema sp PCC 7407
Pseudanabaena sp PCC 7367
Gloeomargarita lithophora strain D10
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
100
2
Acaryochloris marina MBIC11017
Moorea producens JHB
100
Dactylococcopsis salina PCC 8305
Halothece sp PCC 7418
3
93
Cyanothece sp PCC 7424
Pleurocapsa sp PCC 7327
4
100
Microcystis aeruginosa NIES 2549
Microcystis panniformis FACHB 1757
Stanieria cyanosphaera PCC 7437
5
98
100
Cyanothece sp ATCC 51142
Atelocyanobacterium thalassa isolate ALOHA
6
Synechocystis sp PCC 6803
Leptolyngbya sp PCC 7376
Synechococcus sp PCC 8807
79
100
7
99
Geminocystis herdmanii PCC 6308
Geminocystis sp NIES 3708
Cyanobacterium stanieri PCC 7202
Cyanobacterium aponinum PCC 10605
100
8
Planktothrix agardhii NIVA CYA 126/8
77
Trichodesmium erythraeum IMS101
Oscillatoria nigro viridis PCC 7112
Oscillatoria acuminata PCC 6304
Arthrospira platensis NIES 39
Arthrospira sp str PCC 8005
91
100
9
Oscillatoriales cyanobacterium JSC 12
Leptolyngbya boryana dg5
10
Chamaesiphon minutus PCC 6605
Chroococcidiopsis thermalis PCC 7203
Gloeocapsa sp PCC 7428
Crinalium epipsammum PCC 9333
Microcoleus sp PCC 7113
11
Fischerella sp NIES 3754
Rivularia sp PCC 7116
Nostocales cyanobacterium HT 58 2
Anabaena cylindrica PCC 7122
Nostoc punctiforme PCC 73102
Nostoc piscinale CENA21
Nodularia spumigena CCY9414
Anabaena sp 90
Calothrix sp PCC 7507
Cylindrospermum stagnale PCC 7417
Nostoc azollae 0708
Anabaena variabilis ATCC 29413
Nostoc sp PCC 7120
75
99
100
12

Figure 4.5 Species tree based on 16s rRNA gene sequences of 56 species. This species tree contains 12 distinct clades. Colors represent the orders and are same as that of figure 4.1.

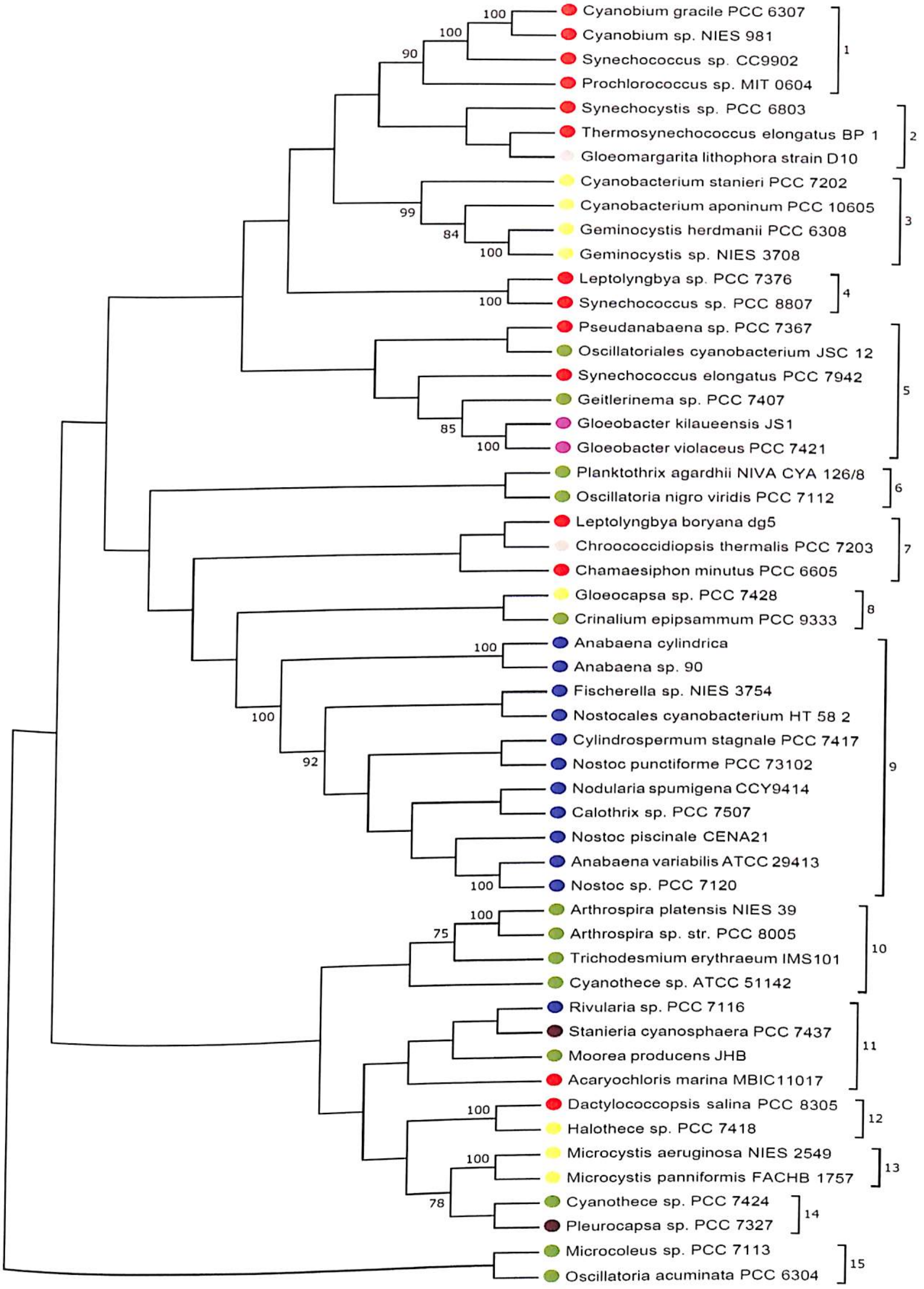

100
100
90
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
Synechococcus sp. CC9902
Prochlorococcus sp. MIT 0604
1
Synechocystis sp. PCC 6803
Thermosynechococcus elongatus BP 1
Gloeomargarita lithophora strain D10
2
99
84
100
Cyanobacterium stanieri PCC 7202
Cyanobacterium aponinum PCC 10605
Geminocystis herdmanii PCC 6308
Geminocystis sp. NIES 3708
3
100
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 8807
4
Pseudanabaena sp. PCC 7367
Oscillatoriales cyanobacterium JSC 12
Synechococcus elongatus PCC 7942
Geitlerinema sp. PCC 7407
85
100
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
5
Planktothrix agardhii NIVA CYA 126/8
Oscillatoria nigro viridis PCC 7112
6
Leptolyngbya boryana dg5
Chroococcidiopsis thermalis PCC 7203
Chamaesiphon minutus PCC 6605
7
Gloeocapsa sp. PCC 7428
Crinalium epipsammum PCC 9333
8
100
Anabaena cylindrica
Anabaena sp. 90
100
92
Fischerella sp. NIES 3754
Nostocales cyanobacterium HT 58 2
Cylindrospermum stagnale PCC 7417
Nostoc punctiforme PCC 73102
Nodularia spumigena CCY9414
Calothrix sp. PCC 7507
Nostoc piscinale CENA21
Anabaena variabilis ATCC 29413
100
Nostoc sp. PCC 7120
9
100
75
Arthrospira platensis NIES 39
Arthrospira sp. str. PCC 8005
Trichodesmium erythraeum IMS101
Cyanothece sp. ATCC 51142
10
Rivularia sp. PCC 7116
Stanieria cyanosphaera PCC 7437
Moorea producens JHB
Acaryochloris marina MBIC11017
11
100
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
12
100
Microcystis aeruginosa NIES 2549
Microcystis panniformis FACHB 1757
13
78
Cyanothece sp. PCC 7424
Pleurocapsa sp. PCC 7327
14
Microcoleus sp. PCC 7113
Oscillatoria acuminata PCC 6304
15

Figure 4.6 Gene tree based on NR gene sequences of 53 species. This tree has 15 distinct clades. Colour coding is the same as figure 4.1.

4.3.1.2.3 Protein tree

Protein tree is more reliable than gene tree as genetic code is degenerate and hence two species which had used different codons for same amino acid could be far in gene tree. Hence we had constructed phylogenetic tree using the available 53 protein sequences (Figure 4.7). The protein tree also showed mixed clade architecture as it was found in the gene tree with species of different orders coming in a single clade. Protein tree was divided into 13 clades. Out of these 13 clades, 8 clades contain species from the same order, i.e. clade 1 and 7 contains species from Synechococales, clade 3 has Chroccocales, clade 4 includes Gloeobacterials, clades 5, 9 and 11 comprises of Oscillatorials and clade 13 includes Nostocales. The remaining 5 clades contains species from different orders. For example, clade 10 has 7 species from 4 orders which are Chrococcales, Oscillatorials, Nostocales and Pleurocapsales. This again points towards a horizontal gene transfer event between the species of these orders. We compared this protein tree with that of the species tree, and we observed that most of the species retained their original positions similar to the species tree, some species followed a different path of evolution, a trend seen in the gene tree analysis. In this case, there are a total of 2 species, *Oscillatoria nigro-viridis* PCC 7112 and *Planktothrix agardhii* NIVA-CYA 126/8, which have changed their association. Both the species with good bootstrap values belong to the order Oscillatorials which reconfirms the diverse pattern of this order as is observed in the species evolution. This suggests that the evolution of these species was influenced by some external environmental pressure which leads to a different evolutionary pattern for this gene in these species. The high level of mixing of species of all orders in protein tree suggests that the cyanobacterial species had extensively communicated with each other at a genetic level. In this tree, *Arthrospira platensis* NIES-39 was present in clade 9 with the same three species as observed in gene tree. Other members of the order Oscillatorials shared different clades with other Orders.

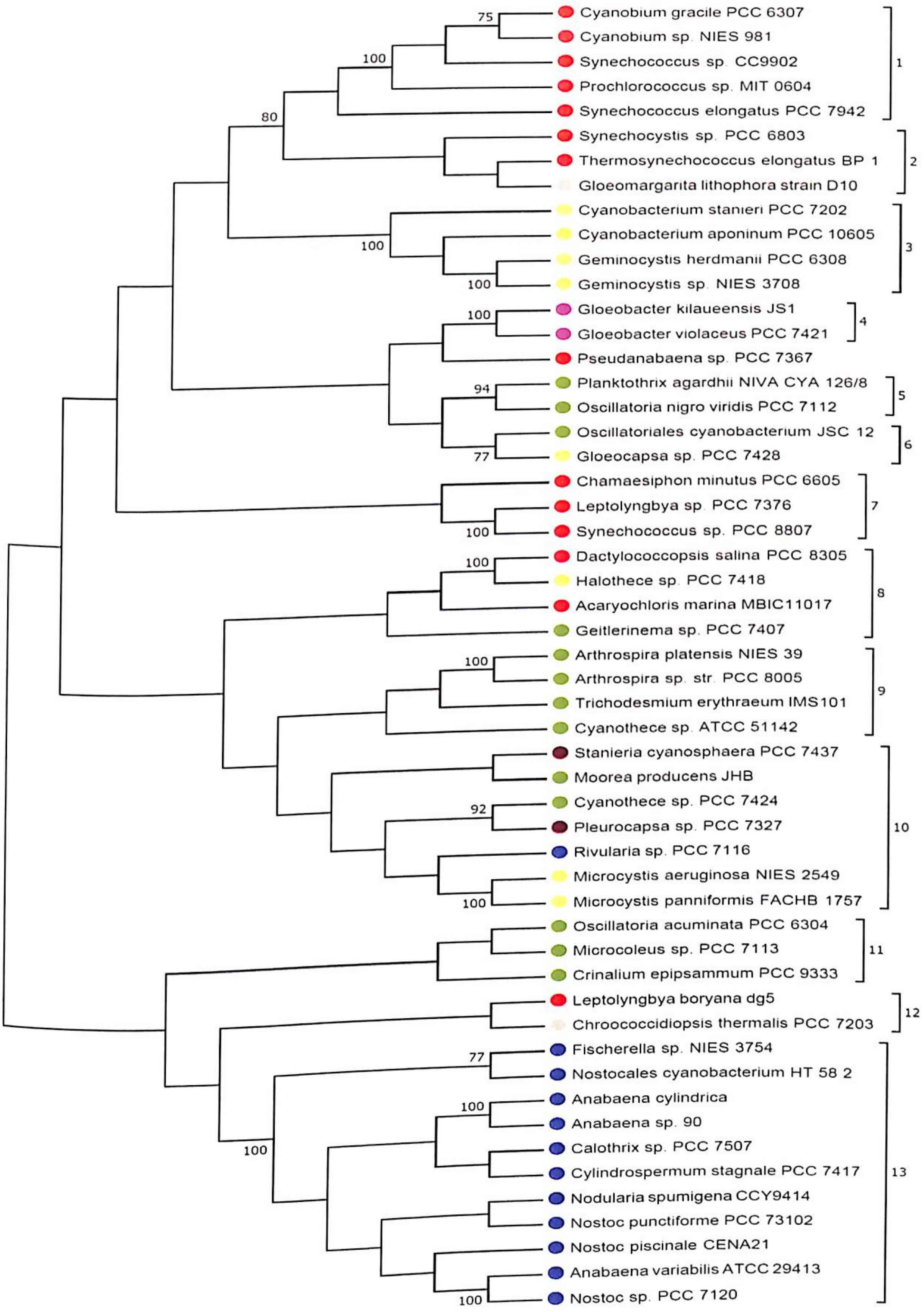

75
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
100
Synechococcus sp. CC9902
Prochlorococcus sp. MIT 0604
80
Synechococcus elongatus PCC 7942
1
Synechocystis sp. PCC 6803
Thermosynechococcus elongatus BP 1
Gloeomargarita lithophora strain D10
2
Cyanobacterium stanieri PCC 7202
100
Cyanobacterium aponinum PCC 10605
Geminocystis herdmanii PCC 6308
100
Geminocystis sp. NIES 3708
3
100
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
4
Pseudanabaena sp. PCC 7367
94
Planktothrix agardhii NIVA CYA 126/8
Oscillatoria nigro viridis PCC 7112
5
Oscillatoriales cyanobacterium JSC 12
77
Gloeocapsa sp. PCC 7428
6
Chamaesiphon minutus PCC 6605
Leptolyngbya sp. PCC 7376
100
Synechococcus sp. PCC 8807
7
100
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Acaryochloris marina MBIC11017
Geitlerinema sp. PCC 7407
8
100
Arthrospira platensis NIES 39
Arthrospira sp. str. PCC 8005
Trichodesmium erythraeum IMS101
Cyanothece sp. ATCC 51142
9
Stanieria cyanosphaera PCC 7437
Moorea producens JHB
92
Cyanothece sp. PCC 7424
Pleurocapsa sp. PCC 7327
Rivularia sp. PCC 7116
Microcystis aeruginosa NIES 2549
100
Microcystis panniformis FACHB 1757
10
Oscillatoria acuminata PCC 6304
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
11
Leptolyngbya boryana dg5
Chroococcidiopsis thermalis PCC 7203
12
77
Fischerella sp. NIES 3754
Nostocales cyanobacterium HT 58 2
100
Anabaena cylindrica
Anabaena sp. 90
100
Calothrix sp. PCC 7507
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc punctiforme PCC 73102
Nostoc piscinale CENA21
Anabaena variabilis ATCC 29413
100
Nostoc sp. PCC 7120
13

Figure 4.7 Protein tree based on NR protein sequences of 53 species. Total 13 distinct clades are observed. Color coding is same as figure 4.1.

4.3.1.3 Codon usages

The codon usages of NR protein in different cyanobacterial species is evaluated by comparing gene and protein trees in a clade wise manner and selected those species which have changed their respective position in the two trees. In gene tree, *Gloeocapsa* belongs to clade 8 with *Crinalium epipsammum* PCC 9333 with a bootstrap value of 31 while in protein tree it was in clade 6 with *Oscillatoriales cyanobacterium* JSC-12 with a bootstrap value of 77. This data suggests that *Gloeocapsa* sp. PCC 7428 and *Oscillatoriales cyanobacterium* JSC-12 may have different codon usage. We have looked into the codon usages of these two species and found that in most cases the two species have used different codons for the same amino acid. To confirm our finding, we created a gene tree based on only first two bases of a codon and leaving the third base. With the third base degeneracy removed in this tree, we observed that *Gloeocapsa* formd a clade with *Oscillatoriales cyanobacterium* JSC-12 with a bootstrap value of 70 (Figure 4.8).

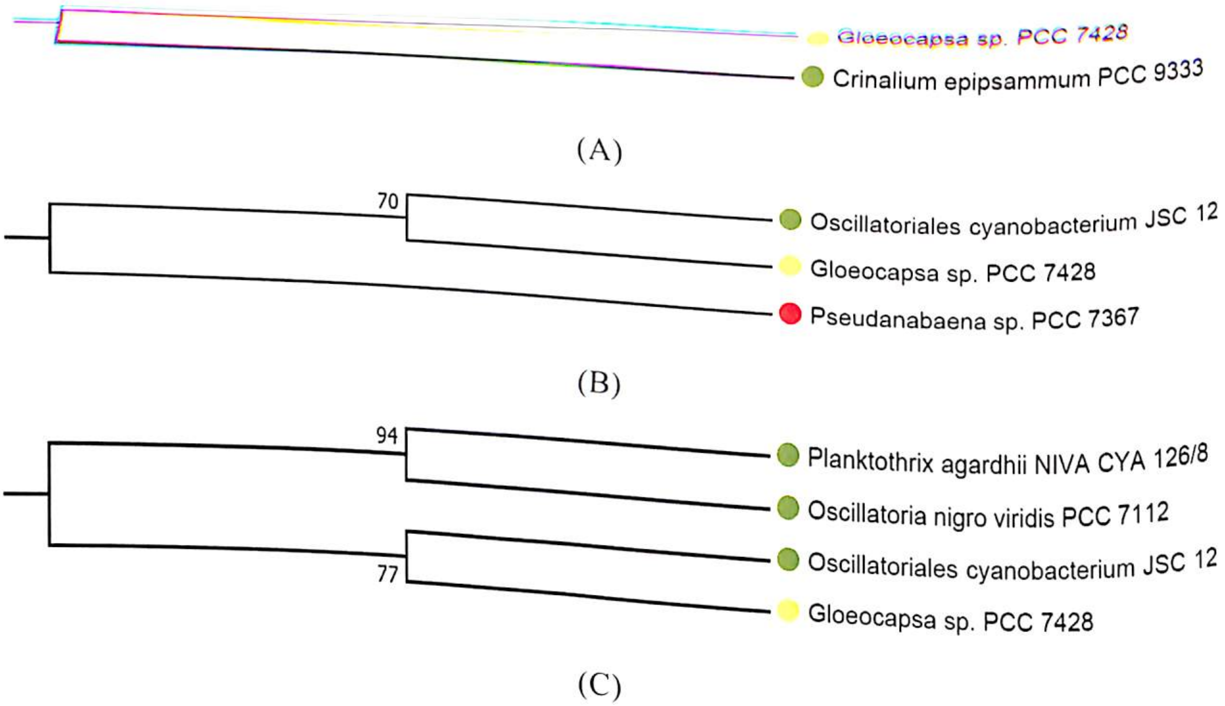

Gloeocapsa sp. PCC 7428
Crinalium epipsammum PCC 9333
(A)
70
Oscillatoriales cyanobacterium JSC 12
Gloeocapsa sp. PCC 7428
Pseudanabaena sp. PCC 7367
(B)
94
Planktothrix agardhii NIVA CYA 126/8
Oscillatoria nigro viridis PCC 7112
77
Oscillatoriales cyanobacterium JSC 12
Gloeocapsa sp. PCC 7428
(C)

Figure 4.8 Codon usages by *Gloeocapsa* sp. PCC 7428 and *Oscillatoriales cyanobacterium* JSC-12. The relative position of *Gloeocapsa* sp. PCC 7428 and *Oscillatoriales cyanobacterium* JSC-12 in (A) NR gene tree (B) NR gene tree based on the first two bases of the codon and (C) NR protein tree.

4.3.1.4 Gene Duplication and Speciation events

Gene duplication and speciation events are the key processes by which genes get transferred from one species to the other. In this study, we used the species tree and the gene tree to examine the possible gene duplication and speciation events in all cyanobacterial species.

The result is depicted within Figure 4.9. The figure indicates that many cyanobacterial species underwent extensive gene duplication and speciation events, which is supported by a good bootstrap value (≥75%). Gene duplication is one of the mechanisms that lead to evolutionary changes. Novel functions or additional functionalities may emerge from the duplicated genes (True and Carroll 2002). Same can be said about speciation events where the same species evolves in different environments and can acquire some additional functionalities to the same proteins.

In case of NR, while the basic functionality was preserved, additional functionality may have come up for this protein in terms of specificity and efficiency. It is evident from the tree (Figure 4.9) that in most cases duplication events happened initially in the evolutionary process which was followed by speciation events. For example, *Geitlerinema* sp. PCC 7407 (soil and freshwater cyanobacteria) and *Gloeobacter* species (found in extreme conditions like limestone and lava caves) which belong to different taxonomic orders but are present in the same clade with high bootstrap support (85%). This gives an idea that both these species have originated from a common speciation event. Similarly, *Dactylococcopsis salina* PCC 8305 and *Halothece* sp. PCC 7418 showed the same behaviour. Speciation events could explain the relatedness of these species of the different orders. These observations further reinforce the widespread diversity of cyanobacterial species arising from different environmental conditions. These duplication and speciation events which lead to the evolution of Nitrate reductase have certainly influenced the functionality of this protein.

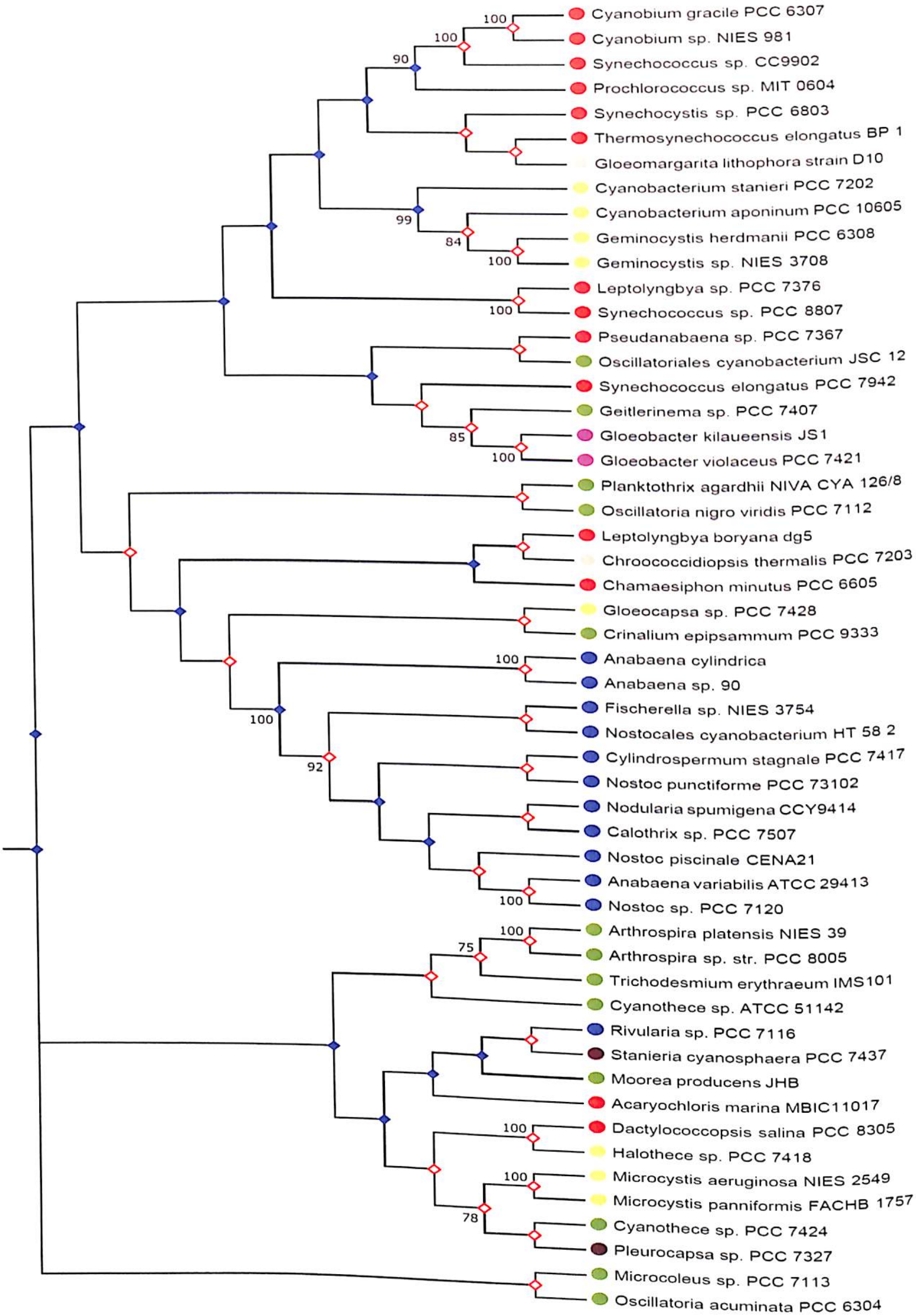

100
Cyanobium gracile PCC 6307
100
Cyanobium sp. NIES 981
90
Synechococcus sp. CC9902
Prochlorococcus sp. MIT 0604
Synechocystis sp. PCC 6803
Thermosynechococcus elongatus BP 1
Gloeomargarita lithophora strain D10
Cyanobacterium stanieri PCC 7202
99
Cyanobacterium aponinum PCC 10605
84
Geminocystis herdmanii PCC 6308
100
Geminocystis sp. NIES 3708
Leptolyngbya sp. PCC 7376
100
Synechococcus sp. PCC 8807
Pseudanabaena sp. PCC 7367
Oscillatoriales cyanobacterium JSC 12
Synechococcus elongatus PCC 7942
Geitlerinema sp. PCC 7407
85
Gloeobacter kilaueensis JS1
100
Gloeobacter violaceus PCC 7421
Planktothrix agardhii NIVA CYA 126/8
Oscillatoria nigro viridis PCC 7112
Leptolyngbya boryana dg5
Chroococcidiopsis thermalis PCC 7203
Chamaesiphon minutus PCC 6605
Gloeocapsa sp. PCC 7428
Crinalium epipsammum PCC 9333
100
Anabaena cylindrica
Anabaena sp. 90
Fischerella sp. NIES 3754
100
Nostocales cyanobacterium HT 58 2
Cylindrospermum stagnale PCC 7417
92
Nostoc punctiforme PCC 73102
Nodularia spumigena CCY9414
Calothrix sp. PCC 7507
Nostoc piscinale CENA21
Anabaena variabilis ATCC 29413
100
Nostoc sp. PCC 7120
100
Arthrospira platensis NIES 39
75
Arthrospira sp. str. PCC 8005
Trichodesmium erythraeum IMS101
Cyanothece sp. ATCC 51142
Rivularia sp. PCC 7116
Stanieria cyanosphaera PCC 7437
Moorea producens JHB
Acaryochloris marina MBIC11017
100
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
100
Microcystis aeruginosa NIES 2549
Microcystis panniformis FACHB 1757
78
Cyanothece sp. PCC 7424
Pleurocapsa sp. PCC 7327
Microcoleus sp. PCC 7113
Oscillatoria acuminata PCC 6304

Figure 4.9 Evolutionary relationships among taxa. There are 3 significant (with bootstrap value >75%) gene duplications (closed diamonds) identified in the tree with 15 significant speciation (open diamonds) events.

4.3.1.5 Structural analysis

To look into the 3-Dimensional structure of the 6 identified residues and the new motif identified in the genus *Arthrospira*, representative species of the clades obtained in the NR protein tree were modelled using the Modeler v9.15 (Table 4.6). A total of eight species were modeled which belonged to four major Orders of the cyanobacteria. The nearest available crystal structure of nitrate reductase was from a bacteria (*Desulfovibrio desulfuricans* ATCC 27774) (PDB-2JIO) which was used as a template. The query coverage was in the range of 95-99%, and the identity was in between 33-37%.

Table 4.6 Three dimentional structures of NR proteins of selected species (based on protein tree) were predicted through homology modelling. Three dimentional structure of NR protein from bacteria *Desulfovibrio desulfuricans* (2JIO) is used as template.

Species modelled	Protein length	Query Coverage (%)	Identity (%)
Arthrospira platensis NIES-39	736	95	35
Dactylococcopsis salina PCC 8305	739	95	36
Geminocystis sp. NIES-3708	715	99	33
Microcystis panniformis FACHB-1757	736	97	37
Nostoc sp. PCC 7120	746	96	36
Planktothrix agardhii NIVA-CYA 126/8	730	96	36
Rivularia sp. PCC 7116	764	95	35
Synechocystis sp. PCC 6803	714	98	36

All the models with validation score analysis are listed in table 4.7. All models were well within the prescribed values for validation. All the modelled structures were superimposed, and we again looked for any variations in each functionally important residue. Our analysis shows that the majority of the functionally important residues are conserved in the cyanobacteria in terms of their orientation in 3-dimensional space. However, there is one region in the protein which showed variation in terms of amino acid composition in various cyanobacteria. This is the region of nitrate entry site, i.e. nitrate enters into the protein and makes its way to the active site. This is a very important site as this is the first point of interaction between the enzyme and the substrate. Since nitrate is a negatively charged molecule, most amino acid residues in this region are positively charged so that they can interact with the substrate more efficiently. This site comprises of nine residues, i.e. S145, R146, M149, A348, R352, A358, N394, C529 and K699 (numbered according to *Arthrospira* *platensis* NIES-39). Out of these nine, S145, R146, M149, A348, R352 and C529 are fully conserved in all the cyanobacterial species while the remaining three residues vary. A358 is replaced with Ser and Cys, Asn394 is replaced with Asp, Thr and Ser while K699 is replaced with Leu and Arg. The variation at position 394 is shown in figure 4.10.

Table 4.7 The quality of predicted NR structure is estimated through various servers and considered to be good structure.

Species	Verify3D	Errat	Q-mean	WhatCheck
Arthrospira platensis NIES-39	84.10	90.634	-3.25	Pass
Dactylococcopsis salina PCC 8305	90.12	91.053	-3.81	Pass
Geminocystis sp. NIES-3708	92.87	90.244	-4.59	Pass
Microcystis panniformis FACHB-1757	88.86	86.060	-4.57	Pass
Nostoc sp. PCC 7120	88.07	91.737	-3.98	Pass
Planktothrix agardhii NIVA-CYA 126/8	92.33	93.285	-3.76	Pass
Rivularia sp. PCC 7116	88.22	87.225	-4.49	Pass
Synechocystis sp. PCC 6803	98.04	94.721	-3.83	Pass

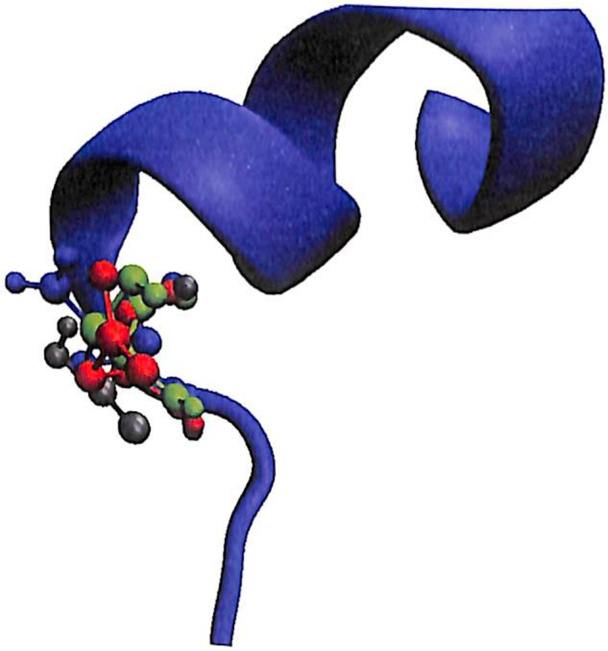


Figure 4.10 Sequence variations, asparagine (blue), aspartic acid (red), threonine (grey) and serine (green) at residue position 394 in 4 superimposed modelled structures is shown. Part of the backbone of *Arthrospira platensis* NIES-39 (blue helix) has been shown.

However, looking deeply into the occurrences of these residues we found out that Asparagine at position 394 is only present in 3 species out of the total 53 species in which NR was found. These species were *Arthrospira platensis* NIES-39, *Arthrospira* sp. str. PCC 8005 and *Chroococcidiopsis thermalis* PCC 7203. This could lead to a different specificity/rate of reaction which in turn will provide more nitrate to the active site and hence more enzyme action.

The three-dimensional structure of the identified motif in *Arthrospira* was determined by *ab-initio* method using Quark server (Xu and Zhang 2012) and has been shown in figure 4.11. A few functionally important residues like K699, L703, K704 and A705 are close to this motif. Hence it might be possible that the presence of this extra motif could affect the functionality of the protein. Also, the C-terminal region of a protein is involved in providing stability to the protein. Hence it is possible that the NR from *Arthrospira platensis* NIES-39 is more stable in terms of half-life and contributes towards the increased protein content.

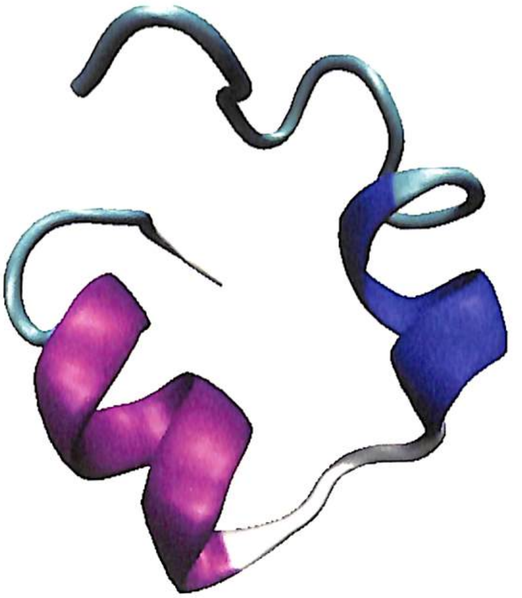


Figure 4.11 Predicted tertiary structure of the C-terminal motif identified in the Genus *Arthrospira*. Two short helices were predicted in this region.

4.3.2 Nitrite reductase (NiR)

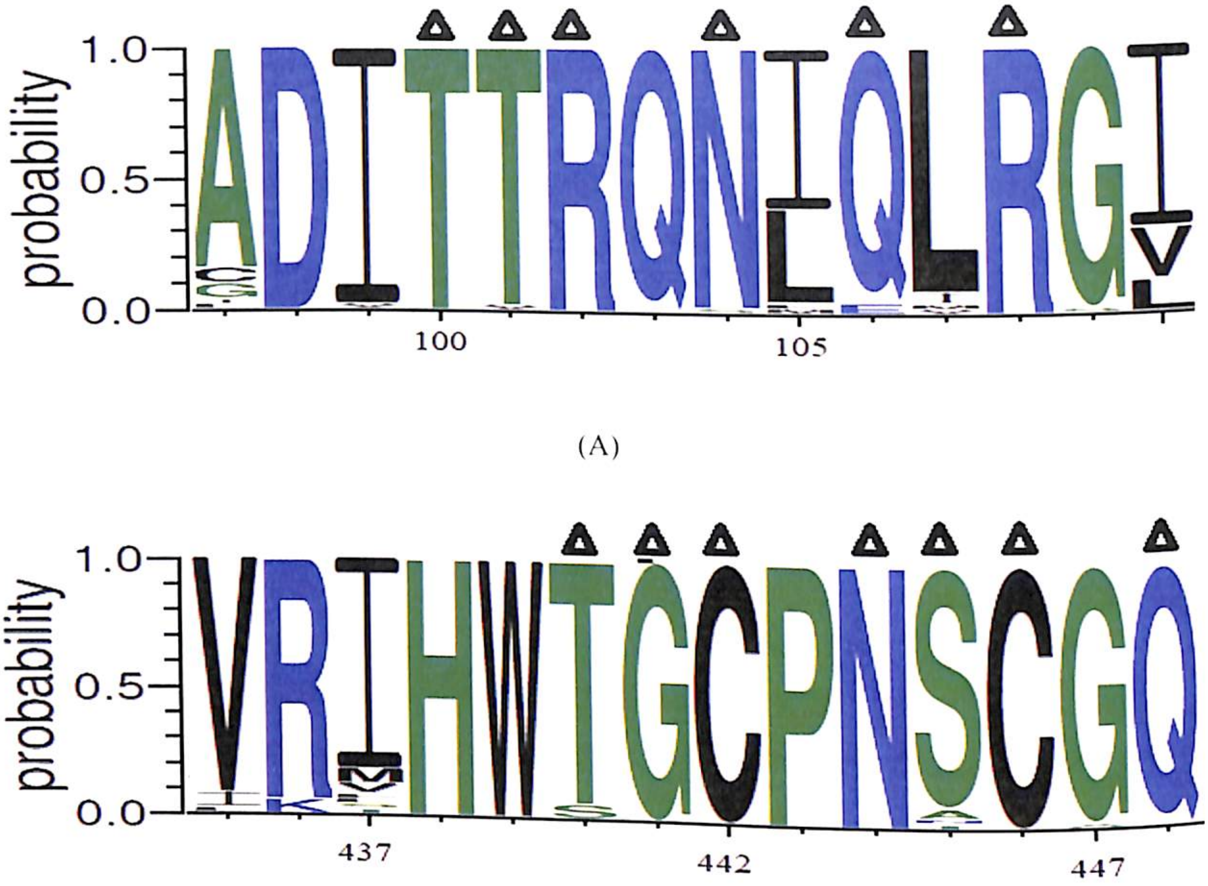
4.3.2.1 Sequence and structural analysis

The average length of this protein was found to be 528. This protein belongs to the Nitrite/Sulfite reductase ferredoxin-like half-domain superfamily. A total of three domains was found in NiR sequences of cyanobacteria, i.e. "NIR_SIR_ferr", "NIR_SIR" and "fer2", the position and length of these domains are listed in Table 4.8. Out of these three domains, the NIR_SIR_ferr and NIR_SIR are present in all the 54 species while fer2 is present only in five cyanobacterial species (*Acaryochloris marina, Leptolyngbya boryana, Crinalium epipsammum, Oscillatoria acuminata* and *Oscillatoria nigro-viridis*). The five cyanobacteria having fer2 extra domain is involved in intramolecular electron transfer to the [4Fe–4S] cluster (Suzuki et al. 1995). We analyzed these domains and reported the signature pattern of these domains in the cyanobacteria. The signature pattern of NIR_SIR_ferr domain has been shown in figure 4.12A. This pattern is 12 amino acids long and contain some functionally important residues like T100, T101, R102, N104, Q106 and R108. The signature pattern for NIR_SIR domain is shown in figure 4.12B. Functionally important residues are T440, G441, C442, N444, S445, C447 and Q448. They are involved in the siroheme binding and Iron-sulphur cluster binding.

Table 4.8 NiR domains position and length. Two major domains were present in this protein. A third domain has been seen in *Acaryochloris marina* MBIC11017, *Leptolyngbya boryana* dg5, *Crinalium epipsammum* PCC 9333, *Oscillatoria acuminata* PCC 6304 and *Oscillatoria nigro-viridis* PCC 7112.

Domains	NIR_SIR_ferr (pfam03460)						NIR_SIR (pfam01077)						fer2 (cd00207)		
Species	From	To	Length	From	To	Length	From	To	Length	From	To	Length	From	To	Length
Acaryochloris marina MBIC11017	53	115	63	303	371	69	123	276	154	381	513	133	557	627	71
Chamaesiphon minutus PCC 6605	61	123	63	311	374	64	131	284	154	385	491	107			
Cyanobium gracile PCC 6307	82	144	63	339	404	66	152	302	151	418	525	108			
Cyanobium sp. NIES-981	70	132	63	324	387	64	140	286	147	403	520	118			
Dactylococcopsis salina PCC 8305	63	124	62	315	380	66	133	288	156	393	503	111			
Leptolyngbya boryana dg5	62	124	63	312	375	64	132	286	155	390	511	122	576	646	71
Leptolyngbya sp. PCC 7376	62	124	63	318	383	66	132	293	162	395	507	113			
Prochlorococcus marinus str. MIT 9313	63	130	68	314	379	66	140	285	146						
Prochlorococcus sp. MIT 0604	57	124	68	310	371	62	132	285	154	387	469	83			
Pseudanabaena sp. PCC 7367	62	124	63	316	381	66	132	291	160	394	501	108			
Synechococcus elongatus PCC 7942	63	125	63	313	378	66	133	288	156	391	501	111			
Synechococcus sp. CC9902	63	130	68	314	379	66	140	291	152	394	503	110			
Synechococcus sp. PCC 8807	64	126	63	319	384	66	134	294	161	396	508	113			
Synechocystis sp. PCC 6803	54	114	61	305	361	57	124	277	154	383	469	87			
Thermosynechococcus elongatus BP-1	54	115	62	305	370	66	124	277	154	383	493	111			
Arthrospira platensis NIES-39	63	125	63	318	383	66	133	293	161	396	503	108			
Arthrospira sp. PCC 8005	63	125	63	318	383	66	133	293	161	396	503	108			
Crinalium epipsammum PCC 9333	58	131	74	319	384	66	139	293	155	397	504	108	575	650	76

Cyanothece sp. ATCC 51142	65	127	63	318	380	63	135	292	158	395	481	87			
Cyanothece sp. PCC 7424	63	123	61	318	383	66	133	293	161	396	506	111			
Geitlerinema sp. PCC 7407	63	123	61	313	369	57	133	287	155	391	501	111			
Microcoleus sp. PCC 7113	63	125	63	314	374	61	133	286	154	391	477	87			
Moorea producens JHB	63	125	63	316	379	64	133	288	156	392	502	111			
Oscillatoria acuminata PCC 6304	66	126	61	316	378	63	134	290	157	393	503	111	548	618	71
Oscillatoria nigro-viridis PCC 7112	65	127	63	315	380	66	135	290	156	393	503	111	546	619	74
Oscillatoriales cyanobacterium JSC-12	63	125	63	313	378	66	133	287	155	390	498	109			
Planktothrix agardhii NIVA-CYA 126/8	63	125	63	313	374	62	133	287	155	391	501	111			
Trichodesmium erythraeum IMS101	63	125	63	313	378	66	133	287	155	391	501	111			
Anabaena cylindrica PCC 7122	63	123	61	312	377	66	133	287	155	389	500	112			
Anabaena sp. 90	63	125	63	318	375	58	133	293	161	395	506	112			
Anabaena variabilis ATCC 29413	64	126	63	331	392	62	134	306	173	405	495	91			
Calothrix sp. PCC 7507	63	125	63	336	406	71	133	311	179	414	500	87			
Cylindrospermum stagnale PCC 7417	64	126	63	319	375	57	134	294	161	393	483	91			
Fischerella sp. NIES-3754	64	120	57	329	394	66	134	304	171	407	491	85			
Nodularia spumigena CCY9414	64	126	63	336	403	68	134	311	178	414	500	87			
Nostoc piscinale CENA21	64	126	63	333	398	66	134	308	175	411	521	111			
Nostoc punctiforme PCC 73102	63	125	63	322	387	66	133	297	165	400	486	87			
Nostoc sp. PCC 7120	64	126	63	331	392	62	134	306	173	405	495	91			
Nostocales cyanobacterium HT-58-2	63	125	63	316	373	58	133	290	158	394	478	85			
Rivularia sp. PCC 7116	65	124	60	316	381	66	135	291	157	394	504	111			
Cyanobacterium aponinum PCC 10605	67	129	63	322	385	64	137	297	161	400	510	111			
Cyanobacterium stanieri PCC 7202	63	125	63	319	382	64	133	294	162	396	507	112			
Geminocystis herdmanii PCC 6308	63	125	63	318	381	64	133	293	161	396	506	111			
Geminocystis sp. NIES-3708	63	125	63	318	381	64	133	293	161	395	503	109			
Gloeocapsa sp. PCC 7428	62	124	63	313	377	65	132	286	155	390	476	87			
Halothece sp. PCC 7418	63	125	63	315	380	66	133	289	157	393	503	111			
Microcystis aeruginosa NIES-2549	64	126	63	316	382	67	134	284	151	393	503	111			
Microcystis panniformis FACHB-1757	64	126	63	316	382	67	134	284	151	393	503	111			
Pleurocapsa sp. PCC 7327	65	127	63	317	377	61	135	290	156	394	504	111			
Stanieria cyanosphaera PCC 7437	62	124	63	315	376	62	132	286	155	391	477	87			
Gloeobacter kilaueensis JS1	48	115	68	304	368	65	123	276	154	380	491	112			
Gloeobacter violaceus PCC 7421	60	127	68	316	376	61	135	288	154	392	503	112			
Chroococcidiopsis thermalis PCC 7203	62	130	69	318	383	66	138	293	156	397	504	108			
Gloeomargarita lithophora Alchichica-D10	54	116	63	308	371	64	124	281	158	385	470	86			



(A)

(B)

Figure 4.12 Conserved sequence patterns for the domains of NiR of all the cyanobacterial sequences. (A) Conserved region of the NIR_SIR_fer domain. The residues having a triangle on top are functionally important residues. T100, T101, R102, N104, Q106 and R108. These residues are involved in the siroheme binding. (B) Conserved region of the NIR_SIR domain. Functionally important residues are T440, G441, C442, N444, S445, C447 and Q448 which are involved in the siroheme binding and Iron-sulphur cluster binding.

Multiple sequence alignment showed the highly conserved nature of the sequences. However, some regions in MSA showed some insertions in few species. To identify any possible new motifs, we used MEME suite program. Due to the highly conserved nature of NiR sequences, most of the motifs were of conserved nature and are present in all the species. 11 major motifs were identified in all the cyanobacteria (Table 4.9).

Table 4.9 MEME output showing the 11 major motifs identified in NiR of the cyanobacterial species. All the identified motifs were highly conserved among cyanobacteria.

Motif	E-Value	Sites	Width
1	2.4e-3207	54	77
2	1.2e-3253	54	77
3	6.4e-3083	54	80
4	1.5e-2879	48	80
5	1.2e-1823	49	44
6	2.1e-1237	53	35
7	2.5e-1087	54	29
8	1.0e-1056	54	29
9	5.1e-342	54	15
10	2.9e-231	48	15
11	1.9e-120	49	8

We have compared cyanobacterial NiR sequence with the tobacco root assimilatory NiR (Nakano et al. 2012), the crystal structure of which is available (PDB ID: 3B0H). Using sequence comparison, we identified 39 functionally important residues in cyanobacteria (Table 4.10). These 39 residues can be divided into 5 categories viz. (1) Seroheme binding – 28 residues (2) Iron Sulphur cluster binding – 5 residues (3) potassium binding – 3 residues (4) Chlorine1 - 2 residues (5) Chlorine2 - 1 residue.

MSA of all cyanobacteria species was analysed for variations in these functionally important residues. Only 4 residues were identified where the amino acid variation could lead to functional efficiency of this NiR. Three positions belonged to the siroheme binding region (226, 270 and 409) while one belonged to the chlorine 1 binding (134) (Numbered according to *Arthropsira platensis* NIES-39). The sequence conservation at these identified 4 positions is given in figure 4.13 in the form of sequence logos.

Table 4.10 Variations found among functionally important residues of NiR in cyanobacteria. Functionally important residues were identified by comparing the sequences of *Nicotiana tabacum* and *Arthrospira platensis* NIES-39.

Nicotiana tabacum	*Arthrospira platensis* NIES-39	*Synechocystis* sp. PCC 6803	Variations in Cyanobacteria	Type of substitution
		Siroheme binding		
F 96	F56	F		
R 98	R58	R		
M 107	M66	M		
R 109	R68	R		

T 141	T100	T		
T 142	T101	T	T(53)V(1)	Differ
R 143	R102	R		
N 145	N104	N	N(53)S(1)	Same
Q 147	Q106	Q	Q(52)E(2)	Differ
R 149	R108	R		
R 223	R182	R		
K 224	K183	K		
N 226	N185	N		
F 264	Y223	F	F(38) Y(10) I(3) V(2)L(1)	Same
F 265	L224	F	F(42) L(12)	Same
S 266	S225	S	S(51) N(2) A(1)	Same
P 267	A226	S	A(41) S(9) P(1) G(3)	Differ
Q 306	Q270	Q	Q(34) N(3) P(2) T(2) L(5) H(1) A(3) G(3)M(1)	Differ
R 309	R273	R		
Q 402	Q363	Q	Q(51) E(3)	Differ
C 441	C401	C		
T 442	T402	T		
K 449	N408	K	N(46) K(2) S(5) G(1)	Differ
	F(409)	F	F(51) L(3)	Same
N 484	N444	N		
S 485	S445	S	S(50) A(2) T(1) N(1)	Same
C 486	C446	C		
Q 488	Q448	Q		
Iron-Sulfur cluster binding				
C 447	C407	C		
A 450	A410	A		
T 480	T440	T	T(51) S(3)	Same
G 481	G441	G		
C 482	C442	C		
Potassium binding				
I 371	V332	V	V(27) I(27)	Same
E 401	E362	E	E(49) D(3) A(2)	Same
N 403	N364	N	N(52) S(2)	Same
Chlorine 1 binding				
M 175	M134	M	M(48)F(2) H(4)	Differ
R 179	R138	R		
Chlorine 2 binding				
R 99	P59	P	P(51)K(3)	Differ
K 100	-	-		
G 448	-	-		

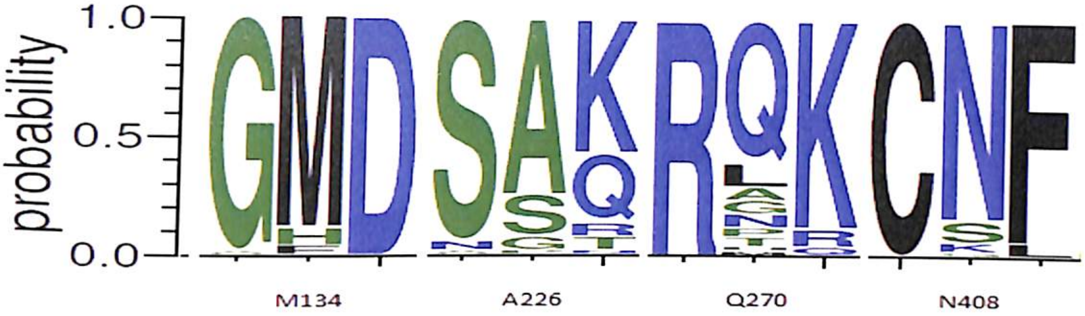


Figure 4.13 Sequence conservations around the functionally important residues (between residue 133-135, 225-227, 269-271 and 407-409) are shown by sequence logo diagram.

4.3.2.2 Phylogenetic analysis

4.3.2.2.1 Gene tree

NiR gene (*nirA*) sequences of 54 NiR species (same as in species tree) were used to generate the NJ phylogenetic tree (Figure 4.14). The gene tree contains 13 distinct clades. Out of these 13 clades, 7 includes species from the same Order, i.e. clade 2, 7 and 13 contain Oscillatorials, clade 3 and 11 have Synechococales, clade 6 includes Chroccocales and clade 9 has Nostocales. The remaining 6 clades contain species from different orders. For example, clade 5 has 4 species from Nostocales, Oscillatorials and Pleurocapsales. We found that the NiR gene from *Rivularia* sp. PCC 7116 (56/26) showed a different evolutionary pattern as was observed in NR. In the species tree, the *Cyanobium* species shared a common ancestor with the *Prochlorococus* species (bootstrap = 100%) which is understandable as both the Genus belong to the same order. But in case of gene tree, *Cyanobium* species shared a common ancestor with *Gloeobacter* species (bootstrap = 100%) which belongs to the order Gloeobacterales. This suggests a high level of genetic interaction between the two Genus/Orders. *Arthrospira platensis* NIES-39 was present in clade 2 of gene tree along with the other species of the Genus *Arthrospira*, i.e. *Arthrospira* sp. str. PCC 8005. Other species of the order Oscillatorials are present in different clades possibly indicating that NiR gene of genus *Arthrospira* could behave differently from other Oscillatorial species.

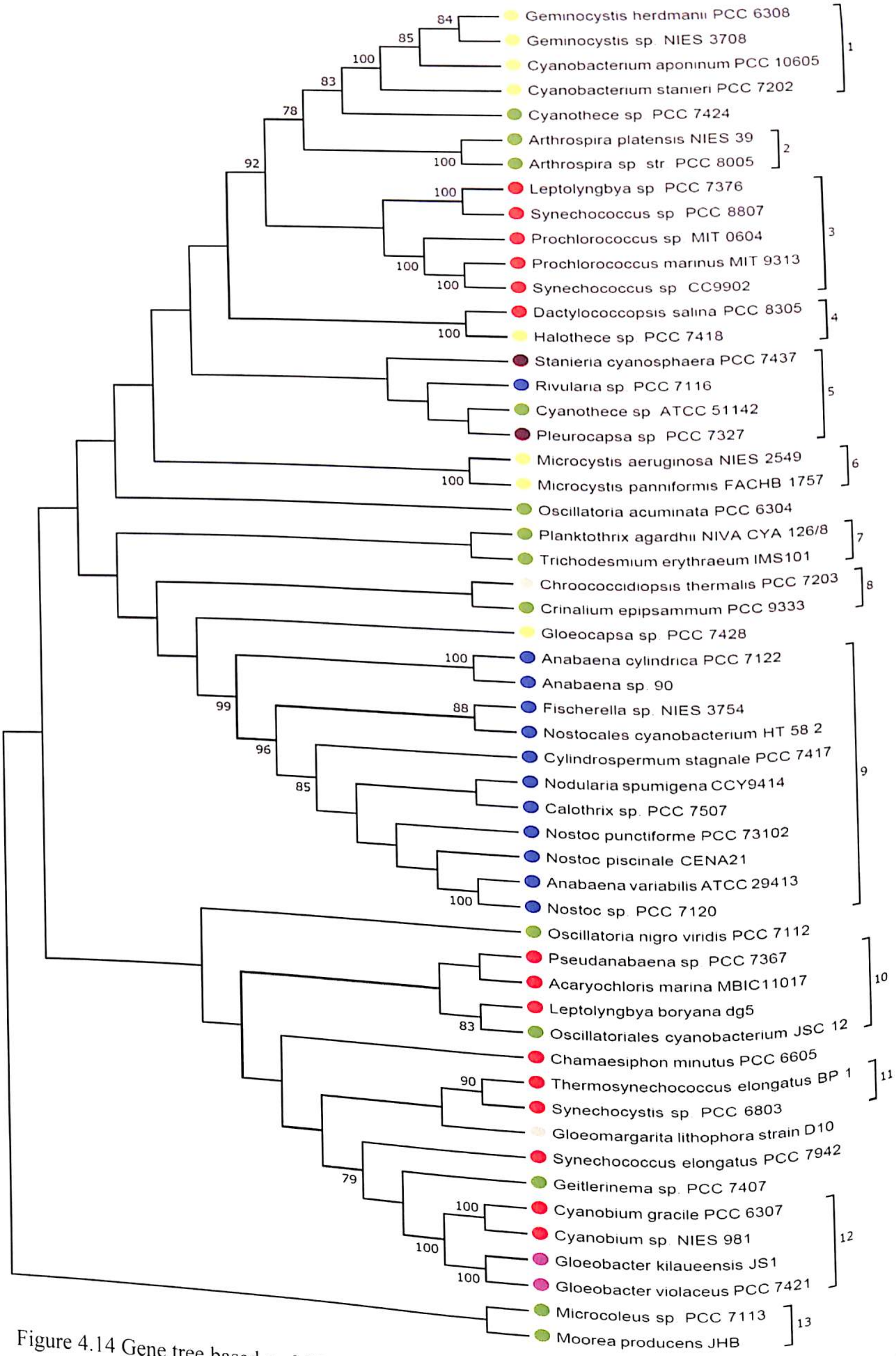

84
Geminocystis herdmanii PCC 6308
85
Geminocystis sp. NIES 3708
100
Cyanobacterium aponinum PCC 10605
83
Cyanobacterium stanieri PCC 7202
1
78
Cyanothece sp. PCC 7424
Arthrospira platensis NIES 39
92
100
Arthrospira sp. str. PCC 8005
2
100
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 8807
Prochlorococcus sp. MIT 0604
3
100
Prochlorococcus marinus MIT 9313
100
Synechococcus sp. CC9902
Dactylococcopsis salina PCC 8305
4
100
Halothece sp. PCC 7418
Stanieria cyanosphaera PCC 7437
Rivularia sp. PCC 7116
5
Cyanothece sp. ATCC 51142
Pleurocapsa sp. PCC 7327
Microcystis aeruginosa NIES 2549
6
100
Microcystis panniformis FACHB 1757
Oscillatoria acuminata PCC 6304
Planktothrix agardhii NIVA CYA 126/8
7
Trichodesmium erythraeum IMS101
Chroococcidiopsis thermalis PCC 7203
8
Crinalium epipsammum PCC 9333
Gloeocapsa sp. PCC 7428
100
Anabaena cylindrica PCC 7122
Anabaena sp. 90
99
88
Fischerella sp. NIES 3754
Nostocales cyanobacterium HT 58 2
96
Cylindrospermum stagnale PCC 7417
9
Nodularia spumigena CCY9414
85
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Nostoc piscinale CENA21
Anabaena variabilis ATCC 29413
100
Nostoc sp. PCC 7120
Oscillatoria nigro viridis PCC 7112
Pseudanabaena sp. PCC 7367
10
Acaryochloris marina MBIC11017
Leptolyngbya boryana dg5
83
Oscillatoriales cyanobacterium JSC 12
Chamaesiphon minutus PCC 6605
90
Thermosynechococcus elongatus BP 1
11
Synechocystis sp. PCC 6803
Gloeomargarita lithophora strain D10
Synechococcus elongatus PCC 7942
79
Geitlerinema sp. PCC 7407
100
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
12
100
Gloeobacter kilaueensis JS1
100
Gloeobacter violaceus PCC 7421
Microcoleus sp. PCC 7113
13
Moorea producens JHB

Figure 4.14 Gene tree based on NiR gene sequences of 54 species contains 13 distinct clades. Color coding is same as figure 4.1.

4.3.2.2.2 Protein tree

Protein phylogenetic tree was made using the 54 protein sequences (Figure 4.15). The protein tree was much more conserved than the gene tree. The protein tree was divided into 14 clades with 11 clades having species from the same order, i.e. Chroccocales in clades 1 and 5, Synechococales in clades 2, 7, 9 and 12, Oscillatorials in clades 3, 10 and 11, Gloeobacterales in clade 8 Nostocales in clade 14. The remaining 3 clades (4, 6 and 13) contained species from different orders. For example, clade 6 contained species from Oscillatorials, Nostocales and Pleurocapsales. A comparison of this protein tree with that of the species tree for differing patterns of evolution did not reveal any species which satisfies the bootstrap cutoff of >75%. However, a general comparison of the topologies of the two trees indicated that the order Oscillatorials is much diverse than other orders for this protein. This is evident as the order Oscillatorial is present in a total 5 clades. This suggests that the NiR gene for this order underwent gene duplication and speciation events. *Arthrospira platensis* NIES-39 shares clade 3 with *Arthrospira* sp. str. PCC 8005 and *Cyanothece* sp. PCC 7424 species of the same order.

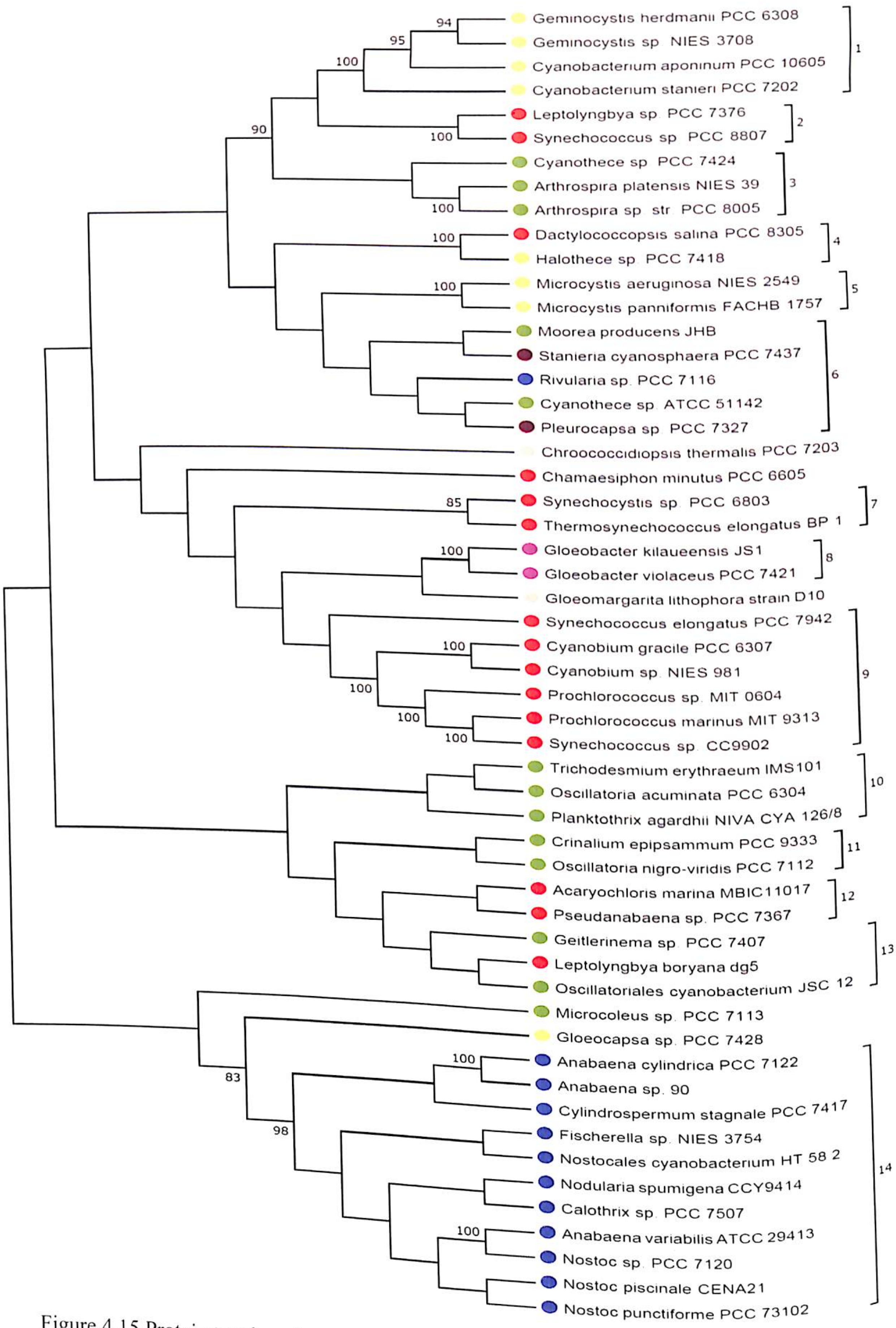
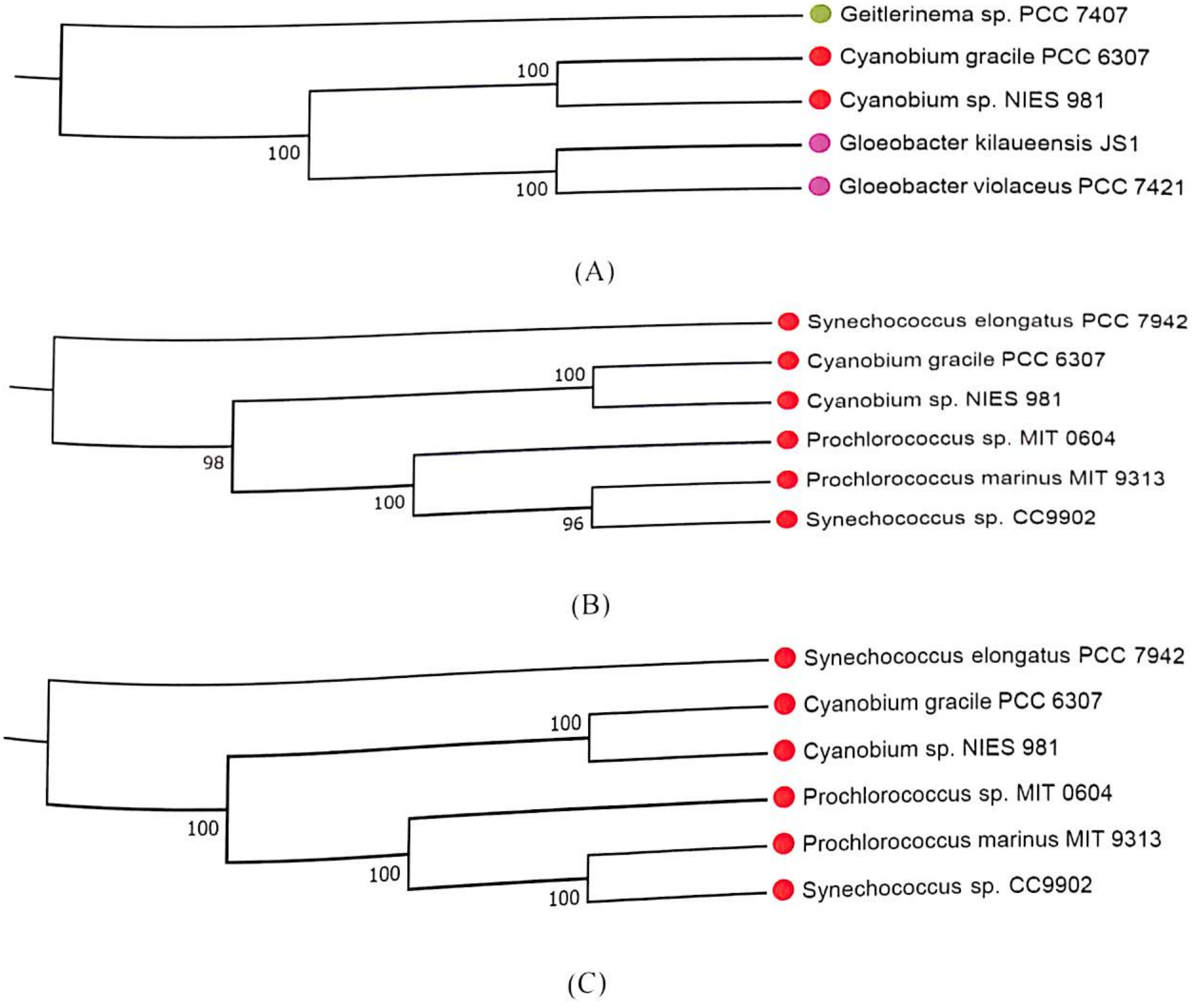

94
95
100
90
100
100
100
100
85
100
100
100
100
100
100
83
98
100
100
Geminocystis herdmanii PCC 6308
Geminocystis sp. NIES 3708
Cyanobacterium aponinum PCC 10605
Cyanobacterium stanieri PCC 7202
1
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 8807
2
Cyanothece sp. PCC 7424
Arthrospira platensis NIES 39
Arthrospira sp. str. PCC 8005
3
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
4
Microcystis aeruginosa NIES 2549
Microcystis panniformis FACHB 1757
5
Moorea producens JHB
Stanieria cyanosphaera PCC 7437
Rivularia sp. PCC 7116
Cyanothece sp. ATCC 51142
Pleurocapsa sp. PCC 7327
6
Chroococcidiopsis thermalis PCC 7203
Chamaesiphon minutus PCC 6605
Synechocystis sp. PCC 6803
Thermosynechococcus elongatus BP 1
7
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
8
Gloeomargarita lithophora strain D10
Synechococcus elongatus PCC 7942
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
Prochlorococcus sp. MIT 0604
Prochlorococcus marinus MIT 9313
Synechococcus sp. CC9902
9
Trichodesmium erythraeum IMS101
Oscillatoria acuminata PCC 6304
Planktothrix agardhii NIVA CYA 126/8
10
Crinalium epipsammum PCC 9333
Oscillatoria nigro-viridis PCC 7112
11
Acaryochloris marina MBIC11017
Pseudanabaena sp. PCC 7367
12
Geitlerinema sp. PCC 7407
Leptolyngbya boryana dg5
Oscillatoriales cyanobacterium JSC 12
13
Microcoleus sp. PCC 7113
Gloeocapsa sp. PCC 7428
Anabaena cylindrica PCC 7122
Anabaena sp. 90
Cylindrospermum stagnale PCC 7417
Fischerella sp. NIES 3754
Nostocales cyanobacterium HT 58 2
Nodularia spumigena CCY9414
Calothrix sp. PCC 7507
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7120
Nostoc piscinale CENA21
Nostoc punctiforme PCC 73102
14

Figure 4.15 Protein tree based on NiR protein sequences of 54 species with 14 distinct clades is shown. Colour coding is the same as figure 4.1.

4.3.2.3 Codon usages

We compared our gene and protein trees to find any evidence of different codon usages. We found that in the gene tree *Cyanobium* species shared a common ancestor with *Gloeobacter* species (bootstrap = 100%) in clade 12 (Figure 4.16A), while in the protein tree *Cyanobium* species shared the common ancestor with the *Prochlorococcus* species (bootstrap = 100%) in clade 9 (Figure 4.16C). We constructed a gene tree based of the first two bases of the codon and found that the position of the *Cyanobium* species was shifted relative to the *Prochlorococcus* species (bootstrap = 98%) (Figure 4.16B). This analysis indicated the importance of the 3rd base in phylogenetic analysis and their role in evolution.


Geitlerinema sp. PCC 7407
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
100
100
100
Synechococcus elongatus PCC 7942
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
Prochlorococcus sp. MIT 0604
Prochlorococcus marinus MIT 9313
Synechococcus sp. CC9902
100
98
100
96
Synechococcus elongatus PCC 7942
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
Prochlorococcus sp. MIT 0604
Prochlorococcus marinus MIT 9313
Synechococcus sp. CC9902
100
100
100
100

(A)

(B)

(C)

Figure 4.16 Codon usages by *Cyanobium* species and the *Prochlorococcus* species. The relative position of *Cyanobium* species and the *Prochlorococcus* species (in (A) NiR gene tree (B) NiR gene tree based on the first two bases of the codon and (C) NiR protein tree.

4.3.2.4 Gene Duplication and Speciation events

In the case of NiR, we found extensive gene duplication and speciation events supported by a good bootstrap value (≥75%) (Figure 4.17). For example, *Leptolyngbya boryana* dg5 (extremophile) and *Oscillatoriales cyanobacterium* JSC 12 (normal fresh water) are present close to each other (bootstrap = 83%) despite belonging to the different orders there by proving their common origin. Similarly, *Dactylococcopsis salina* PCC 8305 and *Halothece* sp. PCC 7418 showed the same behaviour in this protein. These observations have proved the widespread diversity of cyanobacterial species and the effect of evolutionary pressure on the evolution of this protein. However, we have observed that the species that have undergone speciation events in NR and NiR are remarkably similar. This indicates the speciation pattern of both these enzymes, i.e. NR and NiR is similar, and both these enzymes have evolved simultaneously.

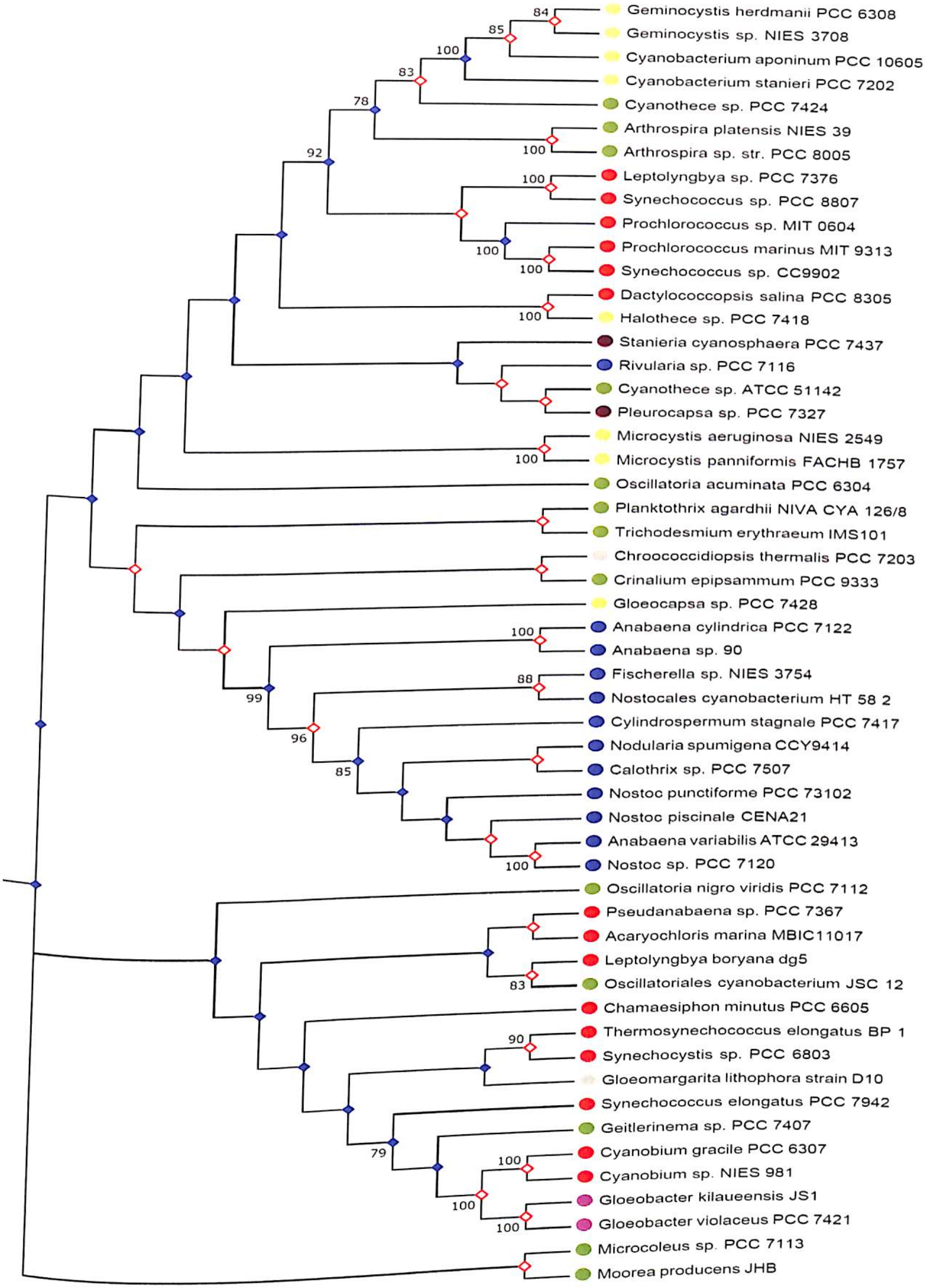

84
85
100
83
78
92
100
100
100
100
100
100
Geminocystis herdmanii PCC 6308
Geminocystis sp. NIES 3708
Cyanobacterium aponinum PCC 10605
Cyanobacterium stanieri PCC 7202
Cyanothece sp. PCC 7424
Arthrospira platensis NIES 39
Arthrospira sp. str. PCC 8005
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 8807
Prochlorococcus sp. MIT 0604
Prochlorococcus marinus MIT 9313
Synechococcus sp. CC9902
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Stanieria cyanosphaera PCC 7437
Rivularia sp. PCC 7116
Cyanothece sp. ATCC 51142
Pleurocapsa sp. PCC 7327
Microcystis aeruginosa NIES 2549
Microcystis panniformis FACHB 1757
Oscillatoria acuminata PCC 6304
Planktothrix agardhii NIVA CYA 126/8
Trichodesmium erythraeum IMS101
Chroococcidiopsis thermalis PCC 7203
Crinalium epipsammum PCC 9333
Gloeocapsa sp. PCC 7428
100
Anabaena cylindrica PCC 7122
Anabaena sp. 90
88
Fischerella sp. NIES 3754
Nostocales cyanobacterium HT 58 2
99
96
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Calothrix sp. PCC 7507
85
Nostoc punctiforme PCC 73102
Nostoc piscinale CENA21
Anabaena variabilis ATCC 29413
100
Nostoc sp. PCC 7120
Oscillatoria nigro viridis PCC 7112
Pseudanabaena sp. PCC 7367
Acaryochloris marina MBIC11017
Leptolyngbya boryana dg5
83
Oscillatoriales cyanobacterium JSC 12
Chamaesiphon minutus PCC 6605
90
Thermosynechococcus elongatus BP 1
Synechocystis sp. PCC 6803
Gloeomargarita lithophora strain D10
Synechococcus elongatus PCC 7942
Geitlerinema sp. PCC 7407
79
100
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
100
Gloeobacter kilaueensis JS1
100
Gloeobacter violaceus PCC 7421
Microcoleus sp. PCC 7113
Moorea producens JHB

Figure 4.17 Evolutionary relationships between taxa. There are 7 significant (with bootstrap value >75%) gene duplications (closed diamonds) identified in the tree with 17 significant speciations (open diamonds) events.

4.3.2.5 Structural analysis

We did homology modelling for 8 representative species of the clades obtained in the NiR protein tree. Two templates were used, namely nitrite reductase from Spinach (PDB ID: 2AKJ), and nitrite reductase from Tobacco root (PDB ID: 3B0H) for 4 species each (Table 4.11). The query coverage for all the species was between 95 to 99% while the identity was between 33-37%. Modelled species covered four major Orders of the cyanobacteria. The results of the validations are shown in Table 4.12.

We superimposed all the modelled structures and looked for any variations in individual functionally important residues with emphasis on the three selected residues A226, Q270 and N408 from *Arthrospira platensis* NIES-39. Our analysis shows that the majority of the functionally important residues are conserved in the cyanobacteria in terms of orientation in the 3-dimensional space. Out of the three, A226 is replaced with Ser, Pro and Gly while Q270 is replaced with Gly, Met, Leu, His, Ala, Thr, Pro and Asp. These replacements are quite similar and are unlikely to affect the protein function.

But analysing the N408 positions, we found out that different residues at this position have different size and structure. Also, the orientation of the two amino acids is quite different (Figure 4.18). Searching the literature, we found out that this kind of substitution has been reported in tobacco (Nakano et al. 2012) where Glutamine (Q) makes the enzyme low-affinity while Lysine (K) makes it high affinity. *Arthrospira platensis* NIES-39 has Asparagine (similar to Glutamine) which makes it low affinity so that it can work rapidly to convert nitrite to ammonium. On the other hand, Lysine makes the enzyme high-affinity and hence works slowly as in *Synechocystic* sp. PCC 6803.

Table 4.11 The three-dimensional structures of NiR protein of selected species (based on clades of protein tree) were predicted through homology modelling. Three dimensional structures of NiR protein from Tobacco (PDB ID-3B0H) and Spinach (PDB Id – 2AKJ) are used as template.

Species	Template used	Organism	Protein	Query Coverage (%)	Identity (%)
Arthrospira platensis NIES-39	3B0H	Tobacco	NiR	97	51
Dactylococcopsis salina PCC 8305	2AKJ	Spinach	NiR	97	50
Geminocystis sp. NIES-3708	3B0H	Tobacco	NiR	97	48
Gloeocapsa sp. PCC 7428	2AKJ	Spinach	NiR	97	52
Nostoc sp. PCC 7120	2AKJ	Spinach	NiR	97	52
Planktothrix agardhii NIVA-CYA 126/8	2AKJ	Spinach	NiR	98	51

Rivularia sp. PCC 7116	3B0H	Tobacco	NiR	97	49
Synechocystis sp. PCC 6803	3B0H	Tobacco	NiR	99	51

Table 4.12 The quality of predicted NiR structure is estimated through various servers and considered to be good structure.

Species	Verify3D	Errat	Q-mean	WhatCheck
Arthrospira platensis NIES-39	89.77	92.531	-4.70	Pass
Dactylococcopsis salina PCC 8305	87.11	90.586	-4.56	Pass
Geminocystis sp. NIES-3708	91.47	94.366	-5.27	Pass
Gloeocapsa sp. PCC 7428	93.60	88.485	-5.30	Pass
Nostoc sp. PCC 7120	93.10	84.095	-5.20	Pass
Planktothrix agardhii NIVA-CYA 126/8	89.63	91.411	-4.25	Pass
Rivularia sp. PCC 7116	87.77	91.075	-5.28	Pass
Synechocystis sp. PCC 6803	86.85	93.644	-3.86	Pass

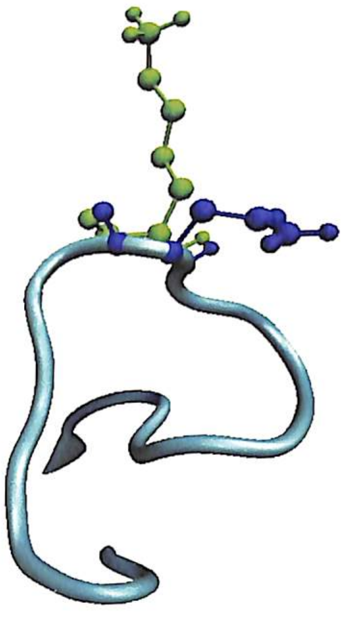


Figure 4.18 Variations at residue position 408 in superimposed modelled structure of *Arthrospira platensis* NIES-39 and *Synechocystis* sp. PCC 6803. Partial backbone structure of only *Arthrospira platensis* NIES-39 (cyan) has been shown. Asparagine (blue) was present in *Arthrospira platensis* NIES-39 while Lysine (green) in *Synechocystis* sp. PCC 6803.

4.4 Conclusions

In this study, Nitrate reductase and Nitrite reductase of *Arthrospira platensis* NIES-39, the two enzymes of the nitrate assimilation pathway were compared within different cyanobacterial species. These enzymes convert nitrate into ammonium which gets into GS/GOGAT pathway and eventually into various nitrogen containing biomolecules. We have compared the sequence and the structural features of these enzymes. The evolutionary process of nitrogen assimilatory enzymes among cyanobacterial species is also examined.

In nitrate reductase, the signature patterns for the domains in cyanobacteria were identified. Sequence analysis identified a 24-amino acid motif (SIVNPELLPTSQTQPNQQQLNPTI) in the Genus *Arthrospira* at the C-terminal. The predicted helical structure of this motif may influence the function of this protein by providing stability to this enzyme as the C-terminal is known to enhance the stability of the protein. Phylogenetic analysis indicates that protein evolution of two Oscillatorials species may be different from that of species evolution. Significant speciation events have been detected in NR. Structural analysis identified a key residue at position 394 in *Arthrospira platensis* NIES-39 sequence. This position is involved in guiding the nitrate towards the active site. It has an Asparagine which is replaced with Serine in *Synechocystis* sp. PCC 6803.

In the case of NiR, we identified the signature patterns of the domains. NiR Gene and protein based phylogenetic analysis revealed the evolutionary process of this gene/protein among cyanobacterial species with probable HGT and speciation events. Structural analysis revealed that the nitrite reductase in cyanobacteria is a dual affinity protein. An amino acid mutation from Asparagine (N) to Lysine (K) at position 408 of *Arthropira platensis* NIES-39 shifts this protein from low affinity to high affinity respectively. This phenomenon was already identified in Tobacco. *Arthrospira platensis* NIES-39 has Asparagine which makes it a low-affinity protein and hence it can work rapidly to convert nitrite to ammonium. This could significantly affects the efficiency of assimilation process.

Chapter V

Comparative analysis of GS-GOGAT pathway enzymes among cyanobacteria

5.1 Introduction

Glutamine synthetase (GS) catalyses the ATP-dependent condensation of glutamate and ammonia to yield glutamine (Liaw et al. 1995). The hydrolysis of ATP drives (Berg et al. 2012) the first step of a two-part concerted mechanism (Liaw et al. 1995, Eisenberg et al. 2000). ATP phosphorylates glutamate to form ADP and an acyl-phosphate intermediate, γ-glutamyl phosphate, which reacts with ammonia, forming glutamine and inorganic phosphate. ADP and P_i do not dissociate until ammonia binds and glutamine is released. Hydrogen bonding and hydrophobic interactions hold the two rings of GS together. Each subunit possesses a C-terminus and an N-terminus in its sequence. The C-terminus (helical thong) stabilises the GS structure by inserting into the hydrophobic region of the subunit across in the other ring. The N-terminus is exposed to the solvent. Also, the central channel is formed via six four-stranded β-sheets composed of anti-parallel loops from the twelve subunits (Eisenberg et al. 2000). The activity of the GSI-type enzyme is controlled by the adenylation of a tyrosine residue. The adenylated enzyme is inactive (Ginsburg et al. 1970).

GOGAT belongs to the family of oxidoreductases, specifically those acting on the $CH-NH_2$ group of donors with an iron-sulfur protein as acceptor. This enzyme synthesises Glutamate from Glutamine and 2-oxoglutarate (2-OG) by transferring the amide group of Glutamine to 2-oxoglutarate resulting in two molecules of Glutamate (Forde and Lea 2007). Cyanobacterial Fd-GOGAT is a monomeric protein while bacterial-GOGAT is a hetero-octamer. Fd-GOGAT and the alpha subunit of NADPH-GOGAT are homologous to each other (Kameya et al. 2007).

This study was aimed at finding the functional role of glutamine synthetase and glutamate synthase in the production of the high protein content of *Arthrospira platensis* NIES-39. Various sequence and structural features were analysed between various cyanobacteria to find any possible explanation for the high protein content of *Arthrospira platensis* NIES-39. We have also considered the evolutionary approach where we compared the 16s based species tree with that of gene/protein tree and looked that whether the gene/protein has evolved in a similar or in a different fashion. Sequence motifs and structural domains across all cyanobacteria were analysed with special emphasis on *Arthrospira platensis* NIES-39. We have also analyzed the functionally important residues of these proteins in *Arthrospira platensis* NIES-39 to look for possible variations that could lead to any functional variation and hence contribute to higher protein content. Structural analyses were also performed to look into any possible structural variations.

5.2 Materials and Methods

5.2.1 Retrieval of Glutamine synthetase and Glutamate synthase protein homologs

Glutamine synthetase and glutamate synthase proteins from *Arthrospira platensis* NIES-39 were used as a query to retrieve the homologs from 56 selected species (as discussed in Chapter IV) from the National Center for Biotechnology Information (NCBI) RefSeq database. Blastn and Blastp (Basic Local Alignment Search Tool) (Altschul et al. 1990) were used against Refseq protein database, and the organism was set to cyanobacteria (taxid 1117) for retrieving the homologous sequences of genes and the proteins from NCBI (with E-value cut off of $\leq 1 \text{ X } 10^{-5}$). We downloaded the GS-I sequences for 54 of the 56 species, and we excluded the remaining two species not possessing a type I GS from our study. Although these two species contain a type III GS (Table 5.1), it is not homologous to type I and hence was excluded from our study. In the case of GOGAT, all the 56 homologs were retrieved. The accession numbers of all the retrieved homologs are given in table 5.1.

Table 5.1 Genome assembly number and the protein accession number for the GS and GOGAT protein of the 56 selected cyanobacteria.

S.No.	Organism Name	Order	Assembly	Protein Accession	
				GS	GOGAT
1	*Acaryochloris marina* MBIC11017	Synechococcales	GCA_000018105.1	WP_012162050.1	WP_041660813.1
2	*Chamaesiphon minutus* PCC 6605	Synechococcales	GCA_000317145.1	Type III	WP_015161811.1
3	*Cyanobium gracile* PCC 6307	Synechococcales	GCA_000316515.1	WP_015108123.1	WP_015109445.1
4	*Cyanobium* sp. NIES-981	Synechococcales	GCA_900088535.1	WP_087067187.1	WP_087068089.1
5	*Dactylococcopsis salina* PCC 8305	Synechococcales	GCA_000317615.1	WP_015229242.1	WP_041235982.1
6	*Leptolyngbya boryana* dg5	Synechococcales	GCA_002142495.1	WP_017287165.1	WP_017288034.1
7	*Leptolyngbya* sp. PCC 7376	Synechococcales	GCA_000316605.1	Type III	WP_015133152.1
8	*Prochlorococcus marinus* str. MIT 9313	Synechococcales	GCA_000011485.1	WP_011129980.1	WP_011131144.1
9	*Prochlorococcus* sp. MIT 0604	Synechococcales	GCA_000757845.1	WP_042850333.1	WP_042851053.1
10	*Pseudanabaena* sp. PCC 7367	Synechococcales	GCA_000317065.1	WP_015164572.1	WP_041698502.1
11	*Synechococcus elongatus* PCC 7942	Synechococcales	GCA_000012525.1	WP_011378345.1	WP_011377778.1
12	*Synechococcus* sp. CC9902	Synechococcales	GCA_000012505.1	WP_011360054.1	WP_041425197.1
13	*Synechococcus* sp. PCC 8807	Synechococcales	GCA_001693295.1	WP_065716519.1	WP_065716969.1
14	*Synechocystis* sp. PCC 6803	Synechococcales	GCA_000340785.1	WP_010871683.1	WP_041426073.1
15	*Thermosynechococcus elongatus* BP-1	Synechococcales	GCA_000011345.1	NP_682378.1	NP_682158.1
16	*Arthrospira platensis* NIES-39	Oscillatorials	GCA_000210375.1	WP_006618330.1	WP_014276035.1

17	*Arthrospira* sp. PCC 8005	Oscillatorials	GCA_000973065.1	CDM94459.1	CDM97111.1
18	*Crinalium epipsammum* PCC 9333	Oscillatorials	GCA_000317495.1	WP_015205278.1	WP_015203962.1
19	*Cyanothece* sp. ATCC 51142	Oscillatorials	GCA_000017845.1	WP_009543512.1	WP_035857095.1
20	*Cyanothece* sp. PCC 7424	Oscillatorials	GCA_000021825.1	WP_015954231.1	WP_012598855.1
21	*Geitlerinema* sp. PCC 7407	Oscillatorials	GCA_000317045.1	WP_015171633.1	WP_041268472.1
22	*Microcoleus* sp. PCC 7113	Oscillatorials	GCA_000317515.1	WP_015181188.1	WP_015185946.1
23	*Moorea producens* JHB	Oscillatorials	GCA_001854205.1	WP_008177966.1	WP_071108092.1
24	*Oscillatoria acuminata* PCC 6304	Oscillatorials	GCA_000317105.1	WP_015148370.1	WP_015150330.1
25	*Oscillatoria nigro-viridis* PCC 7112	Oscillatorials	GCA_000317475.1	WP_015178763.1	WP_015176256.1
26	*Oscillatoriales cyanobacterium* JSC-12	Oscillatorials	GCA_000309945.1	WP_009555714.1	WP_009768838.1
27	*Planktothrix agardhii* NIVA-CYA 126/8	Oscillatorials	GCA_000710505.1	WP_042154436.1	WP_042151381.1
28	*Trichodesmium erythraeum* IMS101	Oscillatorials	GCA_000014265.1	WP_011613207.1	WP_011610318.1
29	*Anabaena cylindrica* PCC 7122	Nostocales	GCA_000317695.1	WP_015212564.1	WP_015214578.1
30	*Anabaena* sp. 90	Nostocales	GCA_000312705.1	WP_015080355.1	WP_015078109.1
31	*Anabaena variabilis* ATCC 29413	Nostocales	GCA_000204075.1	WP_011317041.1	WP_011318130.1
32	*Calothrix* sp. PCC 7507	Nostocales	GCA_000316575.1	WP_015131725.1	WP_015129150.1
33	*Cylindrospermum stagnale* PCC 7417	Nostocales	GCA_000317535.1	WP_015206299.1	WP_015210292.1
34	*Fischerella* sp. NIES-3754	Nostocales	GCA_001548455.1	WP_009453751.1	WP_062242998.1
35	*Nodularia spumigena* CCY9414	Nostocales	GCA_000340565.3	WP_006197273.1	WP_006196392.1
36	*Nostoc azollae* 0708	Nostocales	GCA_000196515.1	WP_013192275.1	WP_013190140.1
37	*Nostoc piscinale* CENA21	Nostocales	GCA_001298445.1	WP_062297509.1	Translated

38	*Nostoc punctiforme* PCC 73102	Nostocales	GCA_000020025.1	WP_012411650.1	WP_012410232.1
39	*Nostoc* sp. PCC 7120	Nostocales	GCA_000009705.1	WP_010996484.1	WP_010998481.1
40	*Nostocales cyanobacterium* HT-58-2	Nostocales	GCA_002163975.1	WP_087541737.1	WP_087542698.1
41	*Rivularia* sp. PCC 7116	Nostocales	GCA_000316665.1	WP_015119315.1	WP_015117220.1
42	*Atelocyanobacterium thalassa* isolate ALOHA	Chroococcales	GCA_000025125.1	WP_012954544.1	WP_012953771.1
43	*Cyanobacterium aponinum* PCC 10605	Chroococcales	GCA_000317675.1	WP_015219678.1	WP_015218362.1
44	*Cyanobacterium stanieri* PCC 7202	Chroococcales	GCA_000317655.1	WP_015224134.1	WP_015222207.1
45	*Geminocystis herdmanii* PCC 6308	Chroococcales	GCA_000332235.1	WP_017296376.1	WP_017293690.1
46	*Geminocystis* sp. NIES-3708	Chroococcales	GCA_001548095.1	WP_066347139.1	WP_066347779.1
47	*Gloeocapsa* sp. PCC 7428	Chroococcales	GCA_000317555.1	WP_015188857.1	WP_041919318.1
48	*Halothece* sp. PCC 7418	Chroococcales	GCA_000317635.1	WP_015226512.1	WP_015225651.1
49	*Microcystis aeruginosa* NIES-2549	Chroococcales	GCA_000981785.1	WP_046660834.1	WP_046662139.1
50	*Microcystis panniformis* FACHB-1757	Chroococcales	GCA_001264245.1	AKV69377.1	AKV65402.1
51	*Pleurocapsa* sp. PCC 7327	Pleurocapsales	GCA_000317025.1	WP_015143979.1	WP_015145655.1
52	*Stanieria cyanosphaera* PCC 7437	Pleurocapsales	GCA_000317575.1	WP_015193159.1	WP_041619963.1
53	*Gloeobacter kilaueensis* JS1	Gloeobacterales	GCA_000484535.1	WP_023172799.1	WP_041243627.1
54	*Gloeobacter violaceus* PCC 7421	Gloeobacterales	GCA_000011385.1	NP_923998.1	NP_924454.1
55	*Chroococcidiopsis thermalis* PCC 7203	Chroococcidiopsidales	GCA_000317125.1	WP_015156652.1	WP_015155462.1
56	*Gloeomargarita lithophora* Alchichica-D10	Gloeoemargaritales	GCA_001870225.1	WP_071454993.1	WP_071454294.1

5.3 Results and Discussions

5.3.1 Glutamine Synthetase (GS)

5.3.1.1 Sequence and structural analysis

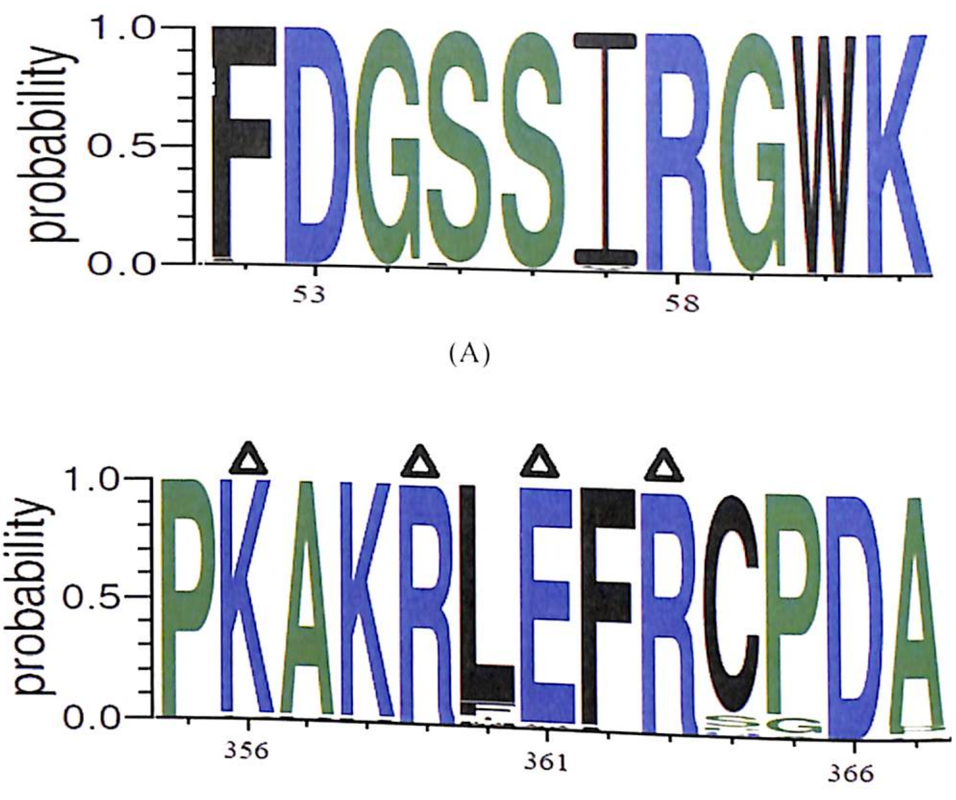
GS of *Arthrospira platensis* NIES-39 is a type I class of GS which is 473 amino acids long homo dodecameric protein (47-56 KDa for a subunit). The average length of the protein was found to be 471.85 amino acids. Two domains were found in GS of all species, i.e. Gln-synt_N (beta grasp domain - pfam03951) and Gln-synt_C (catalytic domain - pfam00120). The position and length of both the domains are listed in Table 5.2. Gln-synt_N domain adopts a beta-grasp fold and contributes to the substrate binding pocket of the enzyme while the catalytic domain helps in the catalysis. We identified the signature patterns of both of the above motifs among cyanobacterial species. For the beta-grasp domain, the signature pattern is 10 amino acids long, that of the catalytic domain contains functionally important residues spanning about 13 amino acids (Figure 5.1).

Except *Cyanobium gracile,* we found that this protein is highly conserved among all the cyanobacterial species. MEME suite of program identified 8 major conserved motifs with significant E-value (Table 5.3). The sequence and modeling analysis indicate that the region between two secondary structure elements is truncated (Figure 5.2 and 5.3).

Table 5.2 Domains boundary of GS protein in each of the cyanobacterial species shows conserved nature of GS protein.

Domains	Gln-synt_N (beta grasp domain) (pfam03951)			Gln-synt_C (catalytic domain) (pfam00120)		
Species	From	To	Length	From	To	Length
Acaryochloris marina MBIC11017	16	97	82	105	470	366
Cyanobium gracile PCC 6307	16	81	66	111	421	311
Cyanobium sp. NIES-981	16	97	82	105	469	365
Dactylococcopsis salina PCC 8305	16	97	82	105	471	367
Leptolyngbya boryana dg5	15	96	82	104	469	366
Prochlorococcus marinus str. MIT 9313	16	97	82	105	470	366
Prochlorococcus sp. MIT 0604	16	97	82	105	470	366
Pseudanabaena sp. PCC 7367	16	97	82	105	470	366
Synechococcus elongatus PCC 7942	16	97	82	105	470	366
Synechococcus sp. CC9902	16	97	82	105	470	366
Synechococcus sp. PCC 8807	16	97	82	105	470	366
Synechocystis sp. PCC 6803	16	97	82	105	470	366
Thermosynechococcus elongatus BP-1	16	95	80	103	468	366
Arthrospira platensis NIES-39	16	97	82	105	470	366
Arthrospira sp. PCC 8005	16	97	82	105	470	366
Crinalium epipsammum PCC 9333	16	97	82	105	470	366

Cyanothece sp. ATCC 51142	16	97	82	105	470	366
Cyanothece sp. PCC 7424	16	97	82	105	470	366
Geitlerinema sp. PCC 7407	16	97	82	105	470	366
Microcoleus sp. PCC 7113	16	97	82	105	470	366
Moorea producens JHB	16	97	82	105	470	366
Oscillatoria acuminata PCC 6304	16	97	82	105	470	366
Oscillatoria nigro-viridis PCC 7112	17	97	81	105	470	366
Oscillatoriales cyanobacterium JSC-12	15	96	82	104	469	366
Planktothrix agardhii NIVA-CYA 126/8	16	97	82	105	470	366
Trichodesmium erythraeum IMS101	16	97	82	105	470	366
Anabaena cylindrica PCC 7122	15	96	82	104	468	365
Anabaena sp. 90	15	96	82	104	468	365
Anabaena variabilis ATCC 29413	15	96	82	104	471	368
Calothrix sp. PCC 7507	15	96	82	104	468	365
Cylindrospermum stagnale PCC 7417	15	96	82	104	468	365
Fischerella sp. NIES-3754	15	96	82	104	468	365
Nodularia spumigena CCY9414	15	96	82	104	468	365
Nostoc azollae 0708	15	96	82	104	474	371
Nostoc piscinale CENA21	15	96	82	104	468	365
Nostoc punctiforme PCC 73102	15	96	82	104	470	367
Nostoc sp. PCC 7120	15	96	82	104	471	368
Nostocales cyanobacterium HT-58-2	15	96	82	104	468	365
Rivularia sp. PCC 7116	15	96	82	104	468	365
Atelocyanobacterium thalassa isolate ALOHA	16	97	82	105	470	366
Cyanobacterium aponinum PCC 10605	16	97	82	105	470	366
Cyanobacterium stanieri PCC 7202	16	97	82	105	470	366
Geminocystis herdmanii PCC 6308	16	97	82	105	470	366
Geminocystis sp. NIES-3708	16	97	82	105	470	366
Gloeocapsa sp. PCC 7428	18	99	82	107	471	365
Halothece sp. PCC 7418	16	97	82	105	471	367
Microcystis aeruginosa NIES-2549	16	97	82	105	470	366
Microcystis panniformis FACHB-1757	16	97	82	105	470	366
Pleurocapsa sp. PCC 7327	16	97	82	105	470	366
Stanieria cyanosphaera PCC 7437	16	97	82	105	470	366
Gloeobacter kilaueensis JS1	16	97	82	105	469	365
Gloeobacter violaceus PCC 7421	16	97	82	105	469	365
Chroococcidiopsis thermalis PCC 7203	16	97	82	105	470	366
Gloeomargarita lithophora Alchichica-D10	16	97	82	105	470	366



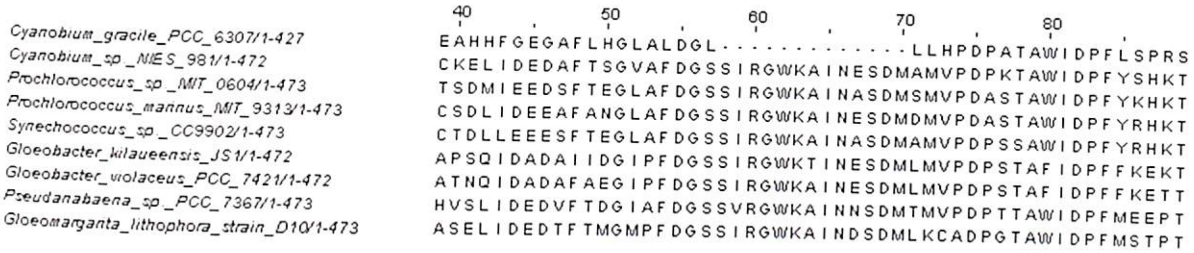
(A)

(B)

Figure 5.1 Signature pattern of the two domains of GS protein (A) beta-grasp domain and (B) catalytic domain. Residues with triangle mark are functionally important.

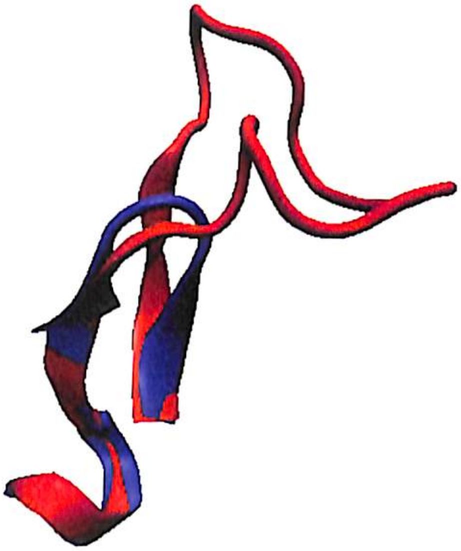
Table 5.3 MEME output showing the statistics for 8 major motifs identified in the GS of cyanobacterial species. Sites represent the number of sequences in which the motif has been identified.

Motif	E-Value	Sites	Width
1	7.0e-3545	37	113
2	3.2e-3319	53	67
3	1.7e-3184	54	63
4	3.3e-4573	53	98
5	1.3e-2053	53	41
6	2.1e-2363	54	57
7	3.4e-919	53	21
8	2.7e-171	42	8



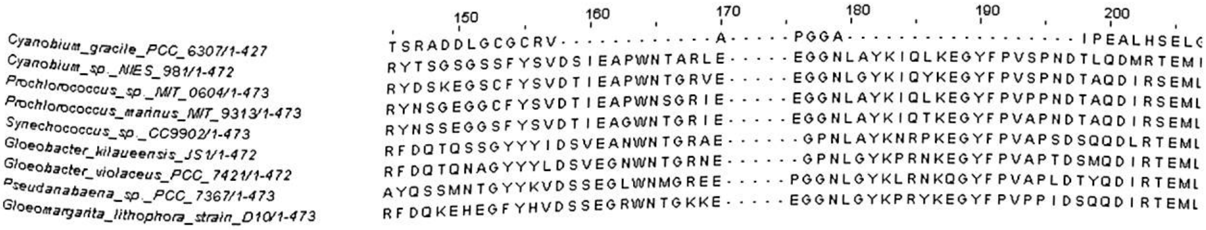
```
                                              40        50        60        70        80
Cyanobium_gracile_PCC_6307/1-427              EAHHFGEGAFLHGLALDGL-------------LLHPDPATAWIDPFLSPRS
Cyanobium_sp._NIES_981/1-472                  CKELIDEDAFTSGVAFDGSSIRGWKAINESDMAMVPDPKTAWIDPFYSHKT
Prochlorococcus_sp._MIT_0604/1-473            TSDMIEEDSFTEGLAFDGSSIRGWKAINASDMSMVPDASTAWIDPFYKHKT
Prochlorococcus_marinus_MIT_9313/1-473        CSDLIDEEAFANGLAFDGSSIRGWKAINESDMDMVPDASTAWIDPFYRHKT
Synechococcus_sp._CC9902/1-473                CTDLLEEESFTEGLAFDGSSIRGWKAINASDMAMVPDPSSAWIDPFYRHKT
Gloeobacter_kilaueensis_JS1/1-472             APSQIDADAIIDGIPFDGSSIRGWKTINESDMLMVPDPSTAFIDPFFKEKT
Gloeobacter_violaceus_PCC_7421/1-472          ATNQIDADAFAEGIPFDGSSIRGWKAINESDMLMVPDPSTAFIDPFFKETT
Pseudanabaena_sp._PCC_7367/1-473              HVSLIDEDVFTDGIAFDGSSVRGWKAINNSDMTMVPDPTTAWIDPFMEEPT
Gloeomargarita_lithophora_strain_D10/1-473    ASELIDEDTFTMGMPFDGSSIRGWKAINDSDMLKCADPGTAWIDPFMSTPT
```

(A)



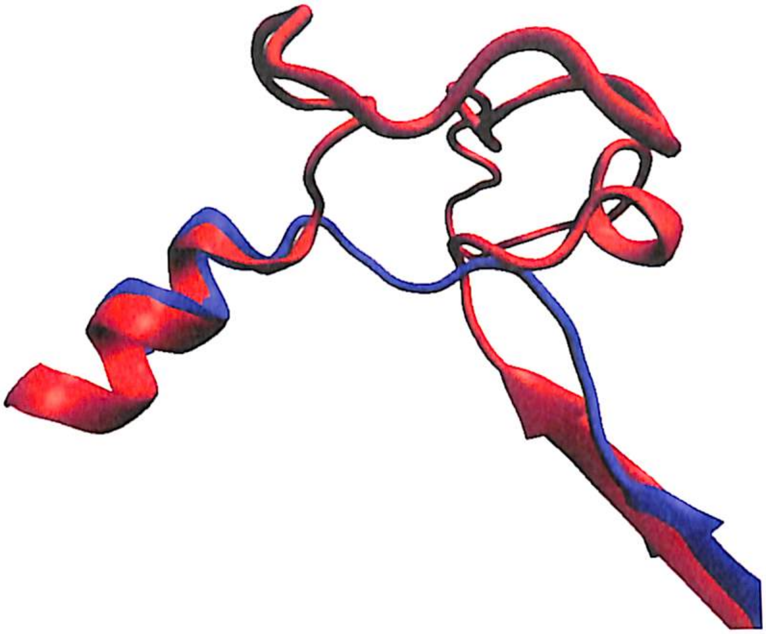
(B)

Figure 5.2(A) Portion of the multiple sequence alignment showing the first deletion in *Cyanobium gracile* (B) Superimposed crystal structure of *Synechocystic* PCC 6803 (red) with modelled *Cyanobium gracile* (blue). The deletion in *Cyanobium* results in a loop deletion.



```
                                                   150       160       170       180       190       200
Cyanobium_gracile_PCC_6307/1-427              TSRADDLGCGCRV------------A-----PGGA------------------IPEALHSELG
Cyanobium_sp._NIES_981/1-472                  RYTSGSGSSFYSVDSIEAPWNTARLE-----EGGNLAYKIQLKEGYFPVSPNDTLQDMRTEMI
Prochlorococcus_sp._MIT_0604/1-473            RYDSKEGSCFYSVDTIEAPWNTGRVE-----EGGNLGYKIQYKEGYFPVSPNDTAQDIRSEML
Prochlorococcus_marinus_MIT_9313/1-473        RYNSGEGGCFYSVDTIEAPWNSGRIE-----EGGNLAYKIQLKEGYFPVPPNDTAQDIRSEML
Synechococcus_sp._CC9902/1-473                RYNSSEGGSFYSVDTIEAGWNTGRIE-----EGGNLAYKIQTKEGYFPVAPNDTAQDIRSEML
Gloeobacter_kilaueensis_JS1/1-472             RFDQTQSSGYYYIDSVEANWNTGRAE------GPNLAYKNRPKEGYFPVAPSDSQQDLRTEML
Gloeobacter_violaceus_PCC_7421/1-472          RFDQTQNAGYYYLDSVEGNWNTGRNE------GPNLGYKPRNKEGYFPVAPTDSMQDIRTEML
Pseudanabaena_sp._PCC_7367/1-473              AYQSSMNTGYYKVDSSEGLWNMGREE-----PGGNLGYKLRNKQGYFPVAPLDTYQDIRTEML
Gloeomargarita_lithophora_strain_D10/1-473    RFDQKEHEGFYHVDSSEGRWNTGKKE-----EGGNLGYKPRYKEGYFPVPPIDSQQDIRTEML
```

(A)



(B)

Figure 5.3(A) Part of the multiple sequence alignment showing the second deletion in *Cyanobium gracile* (B) Superimposed crystal structure of *Synechocystic* PCC 6803 (red) with modelled *Cyanobium gracile* (blue). The deletion in *Cyanobium* results in a deletion of a small helix and loop region.

To identify the functionally important residues in *Arthrospira platensis* NIES-39 that are involved in the activity of this protein, its sequence was compared with *Synechocystis* PCC 6803 whose three-dimensional structure (PDB ID – 3NG0) was available (Saelices et al.). A total of 22 important residues were found conserved (Table 5.4). These 22 residues can be divided into four categories viz. (1) ATP binding – 11 residues (2) Glutamate binding – 5 residues (3) Mn1 binding – 3 residues (4) Mn2 binding – 3 residues.

All the above residues were analysed by looking into the MSA and searching for any type of variation within all the selected cyanobacterial species. Our analysis revealed that the ATP binding Isoleucine226 (Table 5.4) could be substituted with other hydrophobic side chains like Phenylalanine and Valine possibly without any functional alteration.

Table 5.4 Observed sequence variations among functionally important residues of GS in cyanobacteria. Functionally important residues were identified by comparing the sequences of *Synechocystis* sp. PCC 6803 and *Arthrospira platensis* NIES-39.

Synechocystis sp. PCC 6803	*Arthrospira platensis* NIES-39	Variations in Cyanobacteria	Type of substitution
ATP binding residues			
Y 128	F128	F(38) Y(15) R(1)	Same
E 210	E210	E(53)R(1)	Differ
K 211	K211	K(52)L(1)Q(1)	Differ
I 226	I226	F(32) I(15) M(4) L(2) V(1)	Differ
K 227	K227	R(32) K(21) A(1)	Same
F 228	F228	F(53)S(1)	Differ
H 274	H274		
S 276	S276		
R 347	R347	R(53)A(1)	Differ
K 356	K356	K(53)G(1)	Differ
R 359	R359		
Glutamate binding residues			
G 268	G268	G(53)C(1)	Differ
R 324	R324		
E 330	E330	E(53)Q(1)	Differ
R 342	R342	R(53)S(1)	Differ
R 363	R363		
Mn-1 Binding residues			
E132	E132		
H 272	H272	H(53)A(1)	Differ
E 361	E361	E(53)V(1)	Differ
Mn-2 Binding residue			
E 134	E134	E(53)G(1)	Differ
E 215	E215	E(53)Q(1)	Differ
E 223	E223	E(53)H(1)	Differ

5.3.1.2 Phylogenetic analysis

5.3.1.2.1 Gene tree

The gene tree (NJ tree) with 54 *GlnA* (GS gene) contains 11 distinct clades (Figure 5.4). Out of these 11 clades, 5 included species from the same order, i.e. clade 1 contained species from Chroccocales, clade 8 had species from Synechococales, clades 9 and 10 included Oscillatorials and clade 11 contained Nostocales. The other 6 clades contained species from different orders. For example, clade 3 contained 3 species from 3 orders namely Synechococales, Oscillatorials and Chroccocales. Thus, a common evolution of genes has been predicted form this tree. A horizontal gene transfer event could have occured between the cyanobacterial species. To get a clear picture of the gene evolution, we compared the species evolution with the gene evolution. Clade-wise comparison of species and gene tree revealed that while most of the species retained their clades with other co-species in gene tree similar to the species tree, some species moved on to an entirely different clade with different co-species. In these cases, where the species has moved to a new clade, we have considered only those species for which the bootstrap value was high in species as well as in gene tree (>75%). Applying this condition, we found that *Oscillatoria nigroviridis* PCC 7112 had changed its relative position in the gene tree. In the species tree, *Oscillatoria nigroviridis* shared a clade with other species of the order Oscillatorials and particularly with *Trichodesmium erythraeum* with a bootstrap value of 77. However, in the gene tree, it moved away from the rest of the Oscillatorials and makes a clade with another Oscillatorials, i.e. *Crinalium epipsammum* with a bootstrap of 81. This phenomenon hints at the regular exchange of genes at a genetic level between the cyanobacterial species which could be a horizontal gene transfer event. *Arthrospira platensis* NIES-39 was present in clade 9 with other 2 species of the order Oscillatorials which are *Arthrospira* sp. str. PCC 8005 and *Planktothrix agardhii* NIVA CYA. Other members of this order are in different clades in the tree.

5.3.1.2.2 Protein tree

The protein tree (Figure 5.5) was showing much more conserved architecture compared with the gene tree. Protein tree consists of 10 distinct clades. Out of these 10 clades, 7 had species from the same order, i.e. clade 1 had species of Nostocales, clades 2, 3 and 9 included Oscillatorials species, clade 5 contained Gloeobacterials, clade 6 had Synechococales and clade 7 included Chroccocales. The remaining 3 clades contained species from different orders. For example, clade 8 was the most diverse clade and had 7 species from 3 orders namely Chrococcales, Oscillatorials and Synechococales. This mix of clades again points towards a horizontal gene transfer event between the species of these orders. The comparison of the protein tree with the species tree reveals that most of the species retain their original positions as in the species tree, some species have altered their respective clades. In the protein tree, *Arthrospira platensis* NIES-39 was present in clade 2 along with the same species as found in gene tree, i.e. *Arthrospira* sp. str. PCC 8005 and *Planktothrix agardhii* NIVA CYA 128/8. The other species of this order were spread along different clades in the protein tree implying a different evolution of the order Oscillatorials.

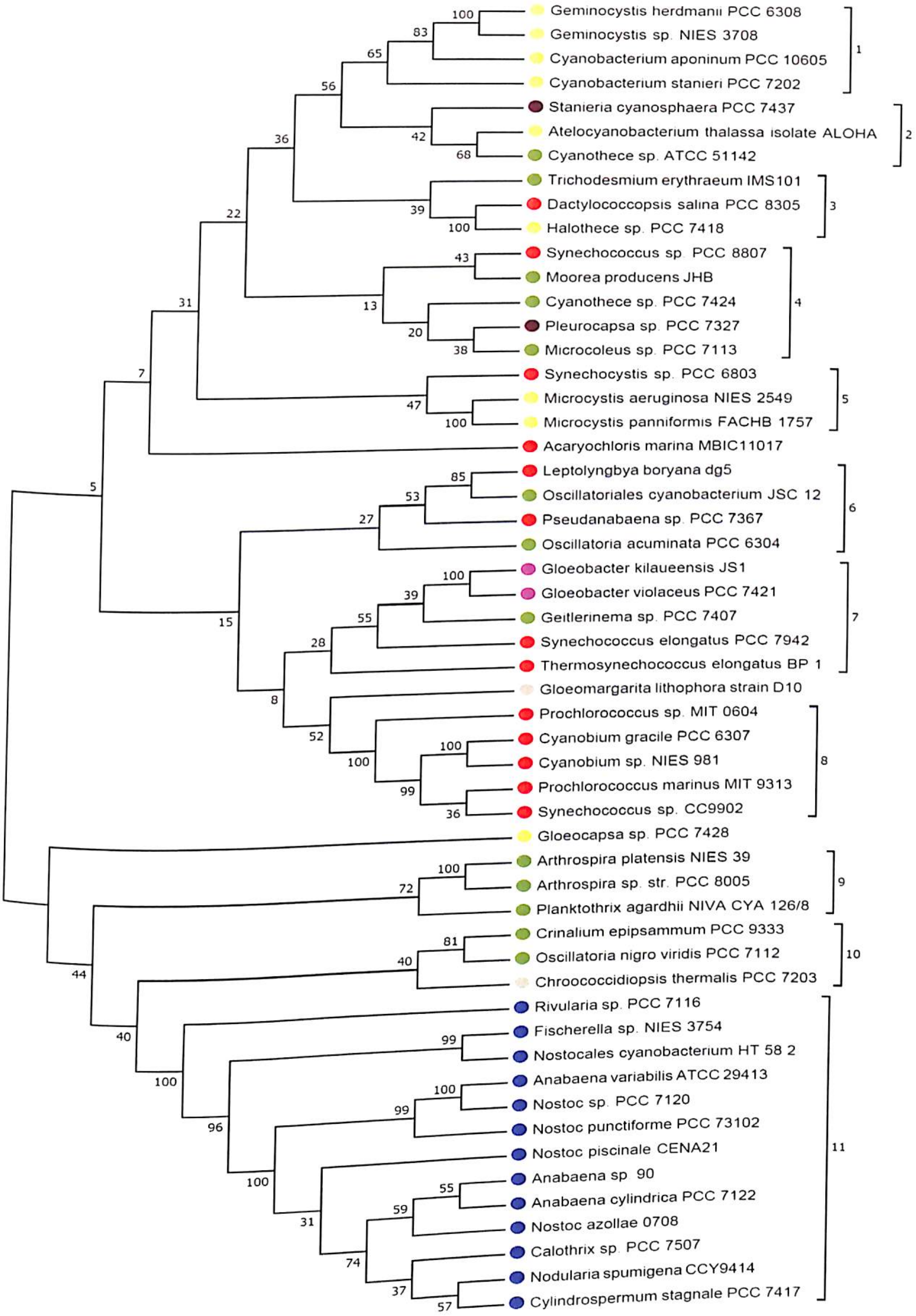

Geminocystis herdmanii PCC 6308
Geminocystis sp. NIES 3708
Cyanobacterium aponinum PCC 10605
Cyanobacterium stanieri PCC 7202
1
Stanieria cyanosphaera PCC 7437
Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
2
Trichodesmium erythraeum IMS101
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
3
Synechococcus sp. PCC 8807
Moorea producens JHB
Cyanothece sp. PCC 7424
Pleurocapsa sp. PCC 7327
Microcoleus sp. PCC 7113
4
Synechocystis sp. PCC 6803
Microcystis aeruginosa NIES 2549
Microcystis panniformis FACHB 1757
5
Acaryochloris marina MBIC11017
Leptolyngbya boryana dg5
Oscillatoriales cyanobacterium JSC 12
Pseudanabaena sp. PCC 7367
Oscillatoria acuminata PCC 6304
6
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Geitlerinema sp. PCC 7407
Synechococcus elongatus PCC 7942
Thermosynechococcus elongatus BP 1
7
Gloeomargarita lithophora strain D10
Prochlorococcus sp. MIT 0604
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
Prochlorococcus marinus MIT 9313
Synechococcus sp. CC9902
8
Gloeocapsa sp. PCC 7428
Arthrospira platensis NIES 39
Arthrospira sp. str. PCC 8005
Planktothrix agardhii NIVA CYA 126/8
9
Crinalium epipsammum PCC 9333
Oscillatoria nigro viridis PCC 7112
Chroococcidiopsis thermalis PCC 7203
10
Rivularia sp. PCC 7116
Fischerella sp. NIES 3754
Nostocales cyanobacterium HT 58 2
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7120
Nostoc punctiforme PCC 73102
Nostoc piscinale CENA21
Anabaena sp 90
Anabaena cylindrica PCC 7122
Nostoc azollae 0708
Calothrix sp. PCC 7507
Nodularia spumigena CCY9414
Cylindrospermum stagnale PCC 7417
11
100
83
65
56
42
68
36
39
100
22
43
13
20
38
31
47
100
7
85
53
27
5
100
39
55
15
28
8
52
100
100
99
36
100
72
81
40
44
40
99
100
100
99
96
100
55
59
31
74
37
57

Figure 5.4 Gene tree based on GS gene sequences of 54 species with 11 distinct clades is shown. Colour coding is the same as figure 4.1.

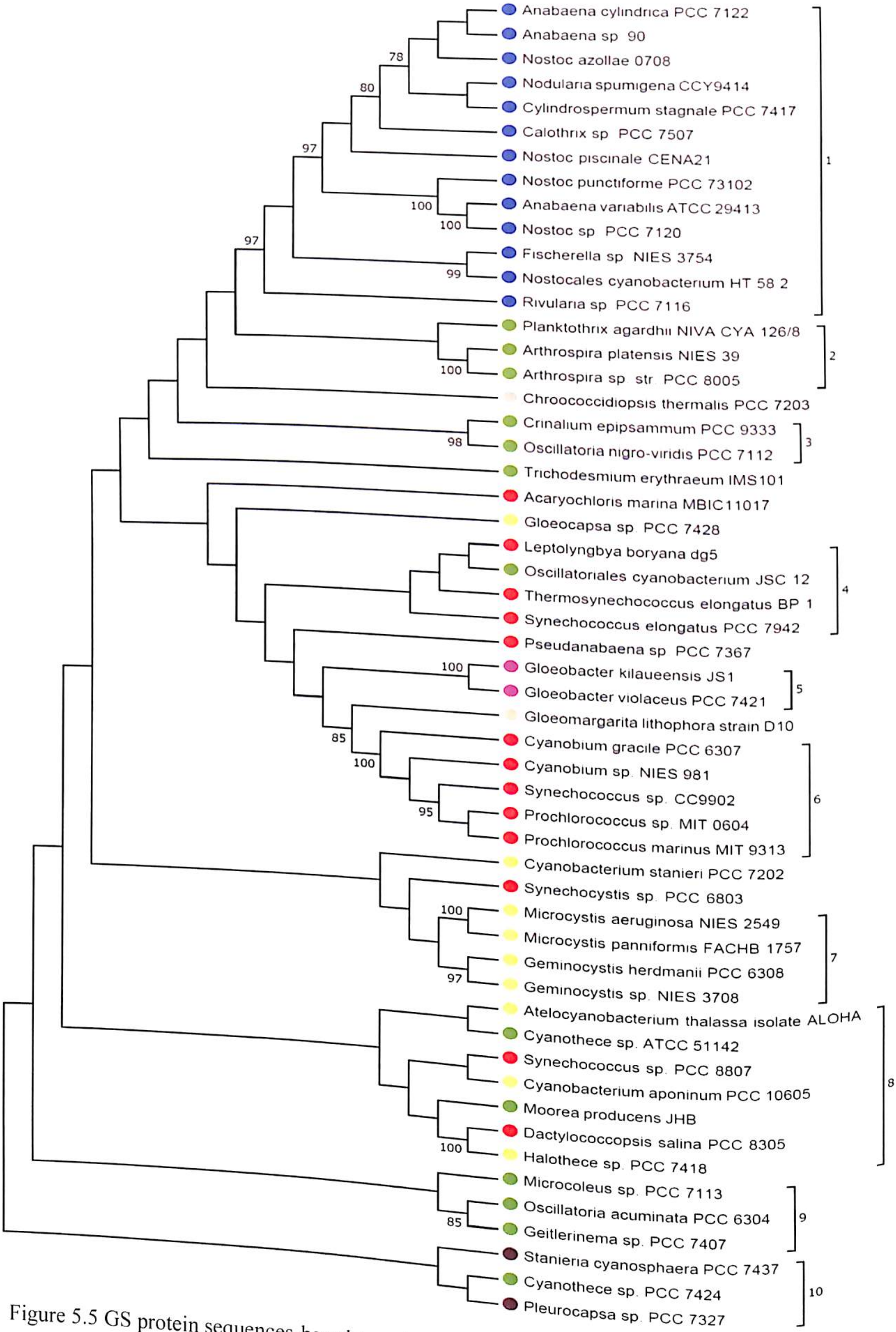

Anabaena cylindrica PCC 7122
Anabaena sp 90
Nostoc azollae 0708
Nodularia spumigena CCY9414
Cylindrospermum stagnale PCC 7417
Calothrix sp PCC 7507
Nostoc piscinale CENA21
Nostoc punctiforme PCC 73102
Anabaena variabilis ATCC 29413
Nostoc sp PCC 7120
Fischerella sp NIES 3754
Nostocales cyanobacterium HT 58 2
Rivularia sp PCC 7116
Planktothrix agardhii NIVA CYA 126/8
Arthrospira platensis NIES 39
Arthrospira sp str PCC 8005
Chroococcidiopsis thermalis PCC 7203
Crinalium epipsammum PCC 9333
Oscillatoria nigro-viridis PCC 7112
Trichodesmium erythraeum IMS101
Acaryochloris marina MBIC11017
Gloeocapsa sp. PCC 7428
Leptolyngbya boryana dg5
Oscillatoriales cyanobacterium JSC 12
Thermosynechococcus elongatus BP 1
Synechococcus elongatus PCC 7942
Pseudanabaena sp PCC 7367
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Gloeomargarita lithophora strain D10
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
Synechococcus sp. CC9902
Prochlorococcus sp. MIT 0604
Prochlorococcus marinus MIT 9313
Cyanobacterium stanieri PCC 7202
Synechocystis sp. PCC 6803
Microcystis aeruginosa NIES 2549
Microcystis panniformis FACHB 1757
Geminocystis herdmanii PCC 6308
Geminocystis sp. NIES 3708
Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Synechococcus sp. PCC 8807
Cyanobacterium aponinum PCC 10605
Moorea producens JHB
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Microcoleus sp. PCC 7113
Oscillatoria acuminata PCC 6304
Geitlerinema sp. PCC 7407
Stanieria cyanosphaera PCC 7437
Cyanothece sp. PCC 7424
Pleurocapsa sp. PCC 7327
78
80
97
100
100
97
99
100
98
100
85
100
95
100
97
100
85
1
2
3
4
5
6
7
8
9
10

Figure 5.5 GS protein sequences-based protein tree with 10 distinct clades is shown. Color coding is same as figure 4.1.

5.3.1.3 Codon usages

It is known that codon degeneracy enables different species to prefer different codons for the same amino acid and hence show variable positions in a gene tree with respect to its corrresponding protein tree. We compared our gene and protein trees in a clade wise manner and selected those species which have changed their respective position in the two trees. We found that *Oscillatoria acuminata* and *Geitlerinema* sp. PCC 7407 were the two species which came close in protein tree while residing in separate clades in the gene tree. Both the species were present in clade 9 in the protein tree (bootstrap=85) but were in clade 6 and 7 respectively in the gene tree. These observations suggested that *Oscillatoria acuminata* and *Geitlerinema sp.* PCC 7407 had used different codons for the same amino acid. We examined the codon usages of these two species and observed that in most cases for the same amino acid the two species have preferred different codons. To confirm our finding, we created a gene tree based on only the first two bases of a codon and not the third base. In this tree, because the degeneracy due to third base had been removed *Oscillatoria acuminata* and *Geitlerinema sp.* PCC 7407 formed a single clade with a bootstrap value of 89 (Figure 5.6).

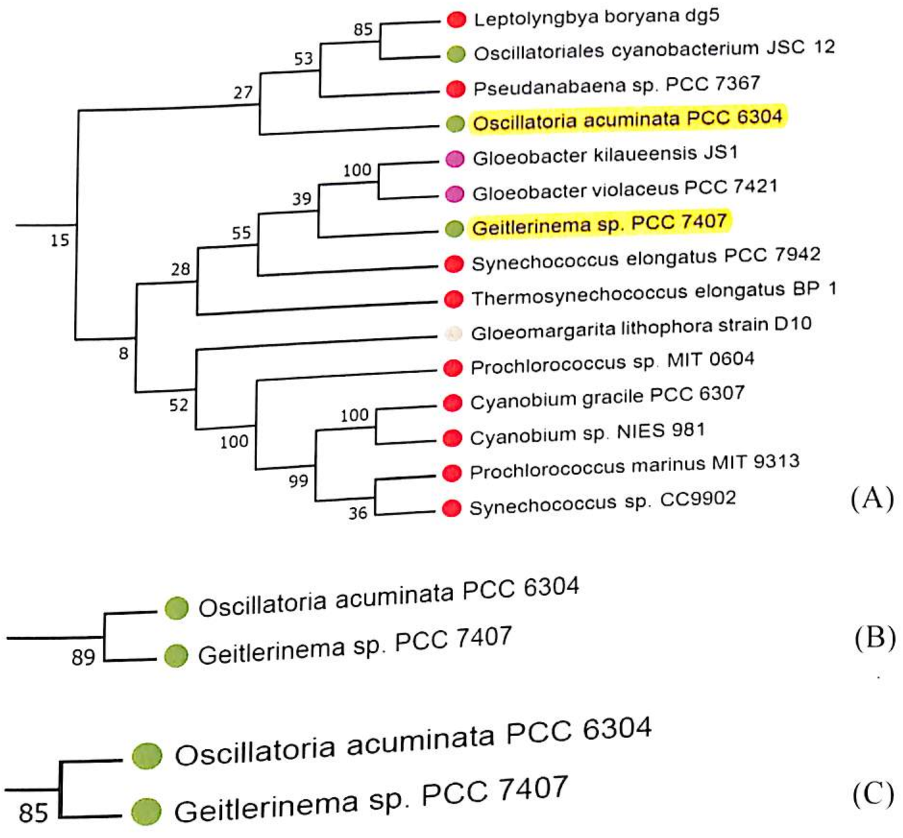

85
53
27
15
100
39
55
28
8
52
100
100
99
36
Leptolyngbya boryana dg5
Oscillatoriales cyanobacterium JSC 12
Pseudanabaena sp. PCC 7367
Oscillatoria acuminata PCC 6304
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Geitlerinema sp. PCC 7407
Synechococcus elongatus PCC 7942
Thermosynechococcus elongatus BP 1
Gloeomargarita lithophora strain D10
Prochlorococcus sp. MIT 0604
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
Prochlorococcus marinus MIT 9313
Synechococcus sp. CC9902
(A)
89
Oscillatoria acuminata PCC 6304
Geitlerinema sp. PCC 7407
(B)
85
Oscillatoria acuminata PCC 6304
Geitlerinema sp. PCC 7407
(C)

Figure 5.6 Codon usages by *Oscillatoria acuminata* and *Geitlerinema sp.* PCC 7407. The relative position of *Oscillatoria acuminata* and *Geitlerinema sp.* PCC 7407 in (A) GS gene tree (B) GS gene tree based on the first two bases of the codon and (C) GS protein tree.

5.3.1.4 Gene Duplication and Speciation events

The GS gene based gene duplication and the speciation events are depicted in Figure 5.7. As observed, a large number of cyanobacterial species has undergone extensive gene duplication and speciation events supported by a good bootstrap value (≥75%). There are three significant gene duplications events (closed diamonds) identified in the tree with 14 significant speciation events (open diamonds). For example, *Leptolyngbya boryana* dg5 (terrestrial and freshwater cyanobacteria) and *Oscillatoriales cyanobacterium* JSC-12 species (found in extreme conditions like hot water springs) belong to different taxonomic orders but are present in the same clade with high bootstrap support (85%). This suggests that these two species originated from a common speciation event. Similarly, *Dactylococcopsis salina* PCC 8305 and *Halothece* sp. PCC 7418 showed the same behaviour. Speciation events could explain the relatedness of these species of the different orders. These observations reinforce the idea that diversification in cyanobacterial species can be driven by differences in environmental conditions. These duplication and speciation events had definitely led to the evolution of Glutamine Synthetase.

5.3.1.5 Structural analysis

GS protein sequences of eight representative species belonging to four major orders of cyanobacteria were modelled using homology modelling technique (Table 5.5). The template used was the crystal structure of Glutamine Synthetase from *Synechocystis* sp. PCC 6803 (PDB ID: 3NG0). The best model (based on N-DOPE score) was energy minimised and was validated using the Verify3D, ERRAT, Qmean score and WhatCheck programs. The results of these validations are shown in Table 5.6.

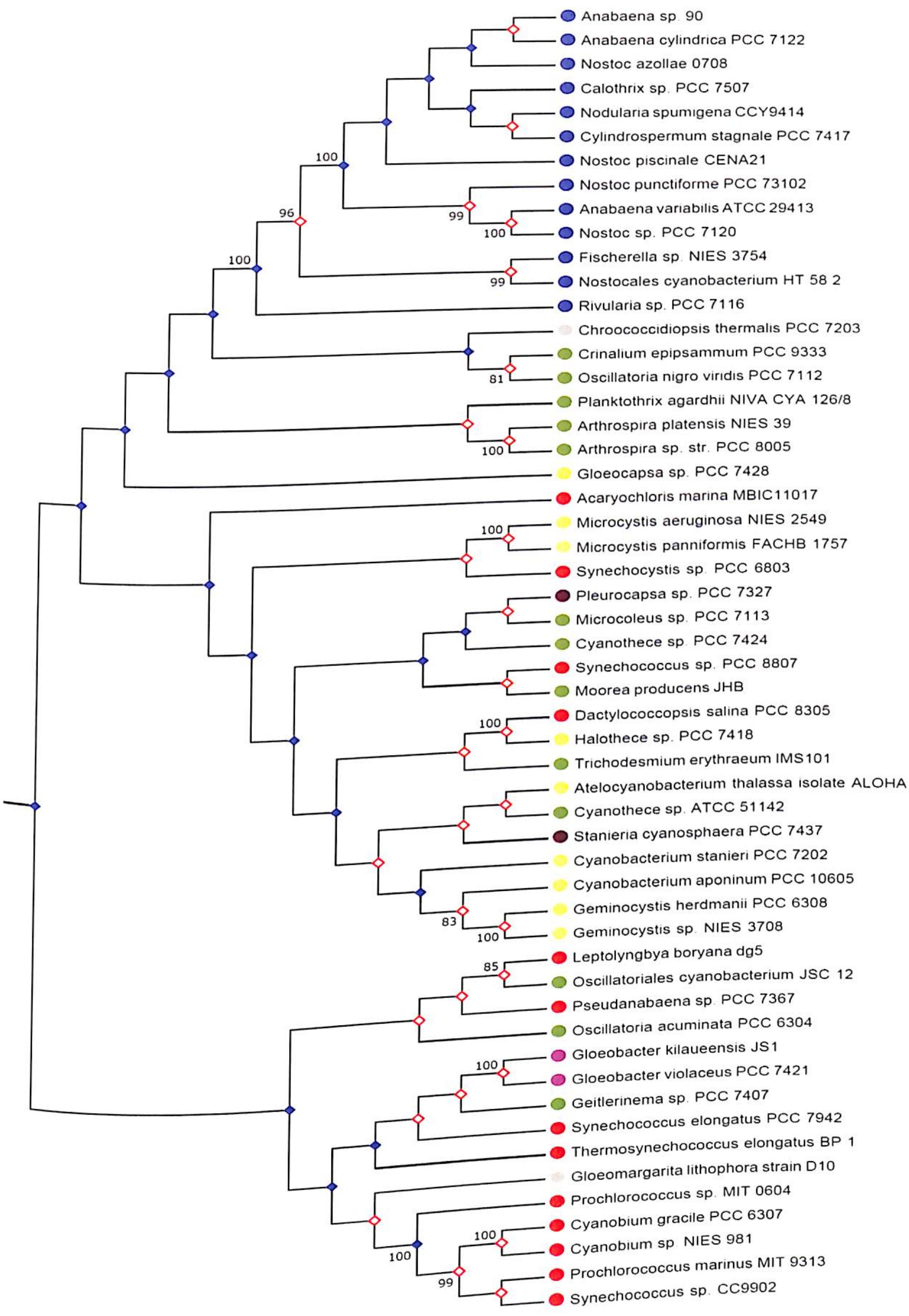

Anabaena sp. 90
Anabaena cylindrica PCC 7122
Nostoc azollae 0708
Calothrix sp. PCC 7507
Nodularia spumigena CCY9414
Cylindrospermum stagnale PCC 7417
Nostoc piscinale CENA21
Nostoc punctiforme PCC 73102
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7120
Fischerella sp. NIES 3754
Nostocales cyanobacterium HT 58 2
Rivularia sp. PCC 7116
Chroococcidiopsis thermalis PCC 7203
Crinalium epipsammum PCC 9333
Oscillatoria nigro viridis PCC 7112
Planktothrix agardhii NIVA CYA 126/8
Arthrospira platensis NIES 39
Arthrospira sp. str. PCC 8005
Gloeocapsa sp. PCC 7428
Acaryochloris marina MBIC11017
Microcystis aeruginosa NIES 2549
Microcystis panniformis FACHB 1757
Synechocystis sp. PCC 6803
Pleurocapsa sp. PCC 7327
Microcoleus sp. PCC 7113
Cyanothece sp. PCC 7424
Synechococcus sp. PCC 8807
Moorea producens JHB
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Trichodesmium erythraeum IMS101
Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Stanieria cyanosphaera PCC 7437
Cyanobacterium stanieri PCC 7202
Cyanobacterium aponinum PCC 10605
Geminocystis herdmanii PCC 6308
Geminocystis sp. NIES 3708
Leptolyngbya boryana dg5
Oscillatoriales cyanobacterium JSC 12
Pseudanabaena sp. PCC 7367
Oscillatoria acuminata PCC 6304
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Geitlerinema sp. PCC 7407
Synechococcus elongatus PCC 7942
Thermosynechococcus elongatus BP 1
Gloeomargarita lithophora strain D10
Prochlorococcus sp. MIT 0604
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
Prochlorococcus marinus MIT 9313
Synechococcus sp. CC9902
100
96
99
100
100
99
81
100
100
100
100
83
100
85
100
100
100
99

Figure 5.7 Evolutionary relationships is shown with 3 significant (with bootstrap value >75%) gene duplication (closed diamonds) events and 14 significant speciation events (open diamonds).

Table 5.5 Three-dimensional structures of GS proteins of selected species (based on protein tree) were predicted through homology modelling. Three-dimensional structure of GS protein of *Synechocystis* sp. PCC 6803 (PDB ID: 3NG0) is considered as a template.

Species	Target length	Query Coverage (%)	Identity (%)
Arthrospira platensis NIES-39	473	100	81
Dactylococcopsis salina PCC 8305	474	100	80
Gloeocapsa sp. PCC 7428	474	99	76
Microcystis panniformis FACHB-1757	473	100	84
Anabaena cylindrica PCC 7122	471	99	80
Cyanothece sp. PCC 7424	473	100	84
Rivularia sp. PCC 7116	471	99	79
Cyanobium gracile PCC 6307	427	98	37

Table 5.6 Validation results for the modelled cyanobacterial species of Glutamine synthetase. All the models were validated using standard validation tools.

Species modelled	Verify3D	Errat	Q-mean	WhatCheck
Arthrospira platensis NIES-39	89.85	96.44	-3.28	Pass
Dactylococcopsis salina PCC 8305	89.03	89.45	-3.45	Pass
Gloeocapsa sp. PCC 7428	82.7	96.86	-2.96	Pass
Microcystis panniformis FACHB-1757	83.51	89.23	-3.70	Pass
Anabaena cylindrica PCC 7122	88.96	95.25	-2.82	Pass
Cyanothece sp. PCC 7424	87.32	95.96	-2.57	Pass
Rivularia sp. PCC 7116	85.56	93.63	-3.10	Pass
Cyanobium gracile PCC 6307	92.27	94.49	-4.03	Pass

All the modelled structures were superimposed and were examined for possible variations at functionally important residues. Our analysis shows that the majority of the functionally important residues are conserved in the cyanobacteria in terms of orientation in the 3-dimensional structures (Figure 5.8).

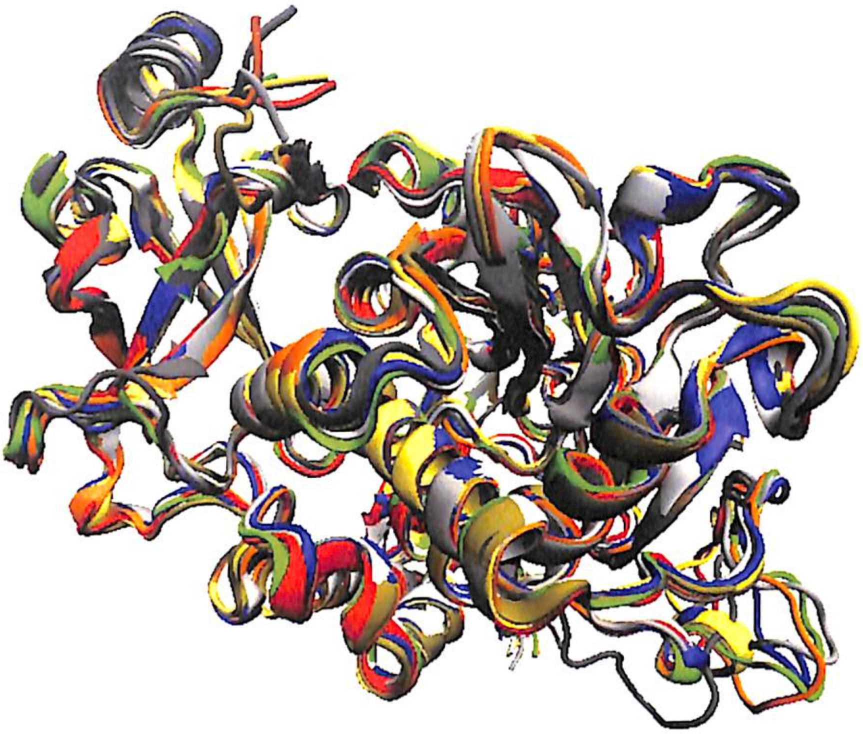


Figure 5.8 Superimposed structures of modelled GS protein of eight species along with the template structure is shown. The colour code is as follows: *Arthrospira platensis* NIES-39 (red), *Dactylococcopsis salina* PCC 8305 (yellow), *Gloeocapsa* sp. PCC 7428 (tan), *Microcystis panniformis* FACHB-1757 (silver), *Anabaena cylindrica* PCC 7122 (blue), *Cyanothece*_sp._PCC_7424 (orange) *Rivularia* sp. PCC 7116 (green), *Cyanobium gracile* PCC 6307 (grey) and *Synechocystic* sp. PCC 6803 (white).

5.3.2 Glutamate Synthase (GOGAT)

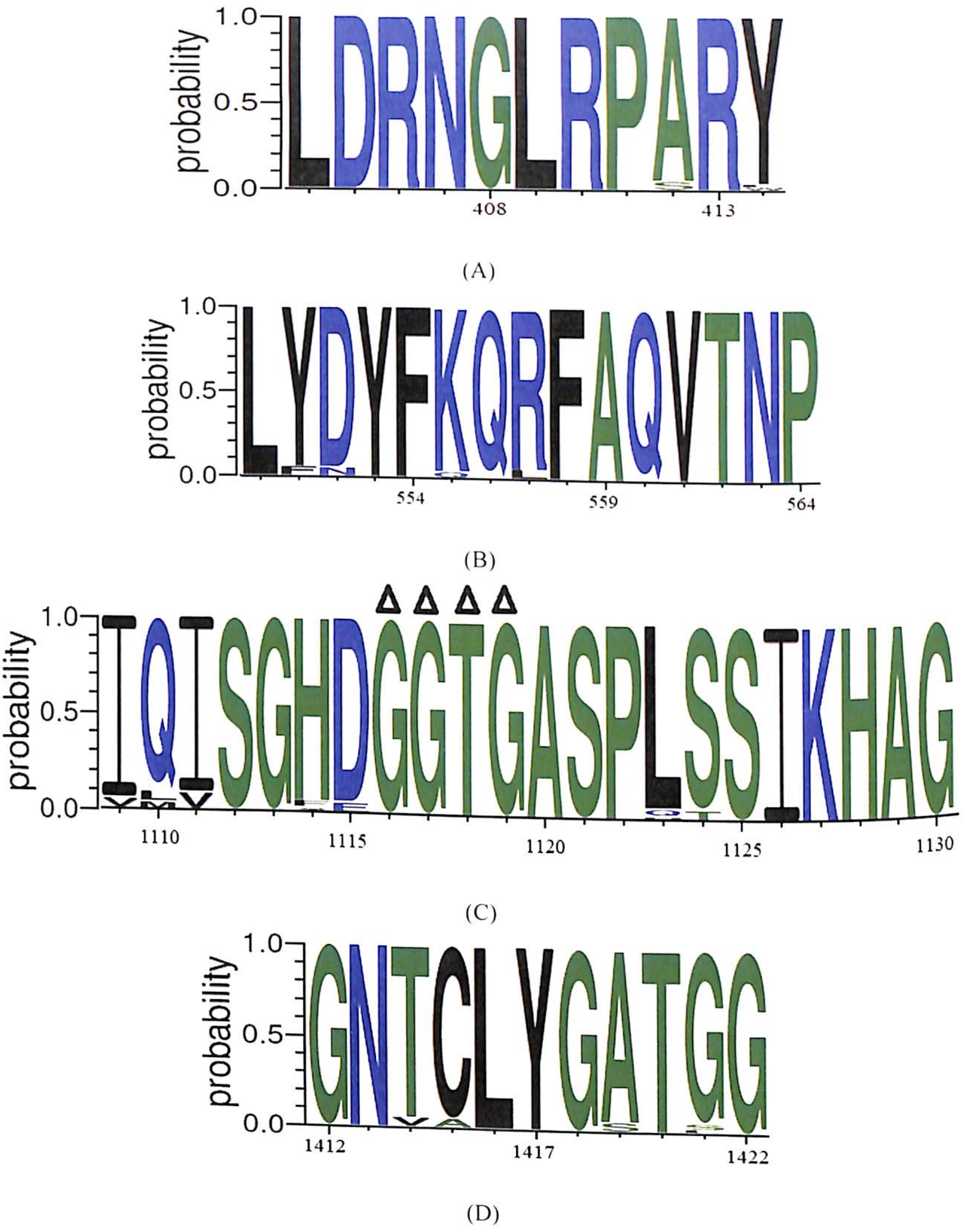
5.3.2.1 Sequence and structural analysis

GOGAT is a large monomeric protein of 1569 amino acids. The average length of the protein in 56 selected species was found to be 1553. Four domains were found in GOGAT of all species, i.e. GATase 2 (pfam00310), Glu_syn_central (pfam04898), Glu_synthase (pfam01645) and a C terminal GXGXG (pfam01493). The position and length of all the domains are listed in Table 5.7. GATase 2 has catalytically important conserved cysteine residue. Central domain connects the amidotransferase domain with the FMN-binding Glu_synthase domain and is highly conserved. FMN binding Glu_synthase domain is the largest domain with a large number of conserved residues. The C terminal GXGXG domain has a mainly structural role in protein function. We identified the signature patterns of these domains. A pattern of 11 residues was identified in GATase 2 domain, a 15-residue long pattern was identified in Glu_syn_central domain, a long pattern of 22 amino acids was identified in the highly conserved Glu_synthase domain and a 11-residue pattern was identified in the GXGXG domain. Sequence conservations of these patterns are shown in figure 5.9 in the form of sequence logos. Around 20 motifs were identified by MEME program. The largest detected motif was of 113 residues.

Table 5.7 GOGAT domains with their respective position and length. Four major domains were present in this protein. All the four domains were present in all the species.

Query	GATase 2 (pfam00310)			Glu_syn_central (pfam04898)			Glu_synthase (pfam01645)			GXGXG (pfam01493)		
	From	To	Length	From	To	Length	From	To	Length	From	To	Length
Acaryochloris marina MBIC11017	25	445	421	471	769	299	827	1212	386	1293	1481	189
Chamaesiphon minutus PCC 6605	37	454	418	481	772	292	829	1218	390	1298	1486	189
Cyanobium gracile PCC 6307	24	449	426	475	768	294	826	1212	387	1293	1477	185
Cyanobium sp. NIES-981	27	445	419	471	761	291	839	1205	367	1286	1473	188
Dactylococcopsis salina PCC 8305	33	456	424	482	775	294	833	1219	387	1299	1487	189
Leptolyngbya boryana dg5	39	459	421	485	783	299	841	1226	386	1306	1494	189
Leptolyngbya sp. PCC 7376	32	449	418	475	768	294	826	1219	394	1299	1487	189
Prochlorococcus marinus str. MIT 9313	28	446	419	474	763	290	821	1207	387	1287	1473	187
Prochlorococcus sp. MIT 0604	28	445	418	471	761	291	819	1205	387	1286	1473	188
Pseudanabaena sp. PCC 7367	32	453	422	479	769	291	827	1213	387	1294	1481	188
Synechococcus elongatus PCC 7942	27	447	421	472	765	294	823	1210	388	1291	1478	188
Synechococcus sp. CC9902	28	443	416	469	767	299	825	1210	386	1293	1474	182
Synechococcus sp. PCC 8807	32	449	418	475	768	294	826	1218	393	1298	1486	189
Synechocystis sp. PCC 6803	27	444	418	471	764	294	821	1217	397	1297	1485	189
Thermosynechococcus elongatus BP-1	28	442	415	468	766	299	824	1210	387	1290	1478	189
Arthrospira platensis NIES-39	45	479	435	506	800	295	858	1244	387	1322	1510	189
Arthrospira sp. PCC 8005	45	479	435	506	800	295	858	1244	387	1322	1510	189

Crinalium epipsammum PCC 9333	35	460	426	486	778	293	836	1222	387	1302	1490	189
Cyanothece sp. ATCC 51142	28	445	418	472	767	296	825	1221	397	1301	1489	189
Cyanothece sp. PCC 7424	32	449	418	476	769	294	827	1223	397	1303	1490	188
Geitlerinema sp. PCC 7407	33	455	423	481	778	298	836	1222	387	1305	1490	186
Microcoleus sp. PCC 7113	35	480	446	506	799	294	857	1243	387	1325	1511	187
Moorea producens JHB	35	507	473	533	826	294	883	1303	421	1383	1570	188
Oscillatoria acuminata PCC 6304	29	459	431	487	775	289	833	1224	392	1305	1493	189
Oscillatoria nigro-viridis PCC 7112	33	456	424	484	784	301	842	1253	412	1333	1521	189
Oscillatoriales cyanobacterium JSC-12	32	458	427	484	782	299	840	1225	386	1306	1494	189
Planktothrix agardhii NIVA-CYA 126 8	22	442	421	470	754	285	815	1185	371	1265	1452	188
Trichodesmium erythraeum IMS101	28	444	417	469	762	294	820	1219	400	1300	1488	189
Anabaena cylindrica PCC 7122	36	450	415	494	795	302	853	1239	387	1319	1507	189
Anabaena sp. 90	36	453	418	502	795	294	853	1239	387	1319	1507	189
Anabaena variabilis ATCC 29413	35	452	418	498	791	294	849	1235	387	1315	1503	189
Calothrix sp. PCC 7507	36	462	427	507	804	298	862	1248	387	1328	1516	189
Cylindrospermum stagnale PCC 7417	36	454	419	505	798	294	856	1242	387	1322	1510	189
Fischerella sp. NIES-3754	36	458	423	506	799	294	857	1243	387	1329	1511	183
Nodularia spumigena CCY9414	36	454	419	492	788	297	846	1232	387	1320	1508	189
Nostoc azollae 0708	36	450	415	494	799	306	857	1243	387	1323	1511	189
Nostoc piscinale CENA21	25	442	418	482	775	294	833	1219	387	1299	1487	189
Nostoc punctiforme PCC 73102	35	453	419	496	789	294	847	1233	387	1313	1501	189
Nostoc sp. PCC 7120	35	452	418	495	788	294	846	1232	387	1312	1500	189
Nostocales cyanobacterium HT-58-2	36	458	423	500	793	294	851	1237	387	1317	1505	189
Rivularia sp. PCC 7116	31	453	423	563	865	303	923	1309	387	1389	1577	189
Atelocyanobacterium thalassa isolate ALOHA	13	429	417	457	751	295	809	1204	396	1285	1472	188
Cyanobacterium aponinum PCC 10605	22	442	421	472	776	305	836	1206	371	1286	1474	189
Cyanobacterium stanieri PCC 7202	25	443	419	476	783	308	841	1231	391	1311	1499	189
Geminocystis herdmanii PCC 6308	31	449	419	485	822	338	880	1270	391	1350	1538	189
Geminocystis sp. NIES-3708	30	448	419	477	767	291	825	1215	391	1295	1483	189
Gloeocapsa sp. PCC 7428	35	458	424	507	799	293	857	1244	388	1326	1512	187
Halothece sp. PCC 7418	40	463	424	489	782	294	840	1226	387	1306	1494	189
Microcystis aeruginosa NIES-2549	31	447	417	472	763	292	821	1210	390	1291	1477	187
Microcystis panniformis FACHB-1757	22	440	419	467	749	283	809	1179	371	1259	1444	186
Pleurocapsa sp. PCC 7327	32	449	418	476	770	295	828	1226	399	1306	1494	189
Stanieria cyanosphaera PCC 7437	34	451	418	479	770	292	828	1214	387	1294	1482	189
Gloeobacter kilaueensis JS1	27	451	425	478	769	292	827	1219	393	1298	1472	175
Gloeobacter violaceus PCC 7421	18	442	425	469	760	292	818	1210	393	1289	1471	183
Chroococcidiopsis thermalis PCC 7203	35	457	423	506	799	294	857	1243	387	1323	1510	188
Gloeomargarita lithophora Alchichica-D10	16	434	419	460	752	293	810	1196	387	1275	1462	188



(A)

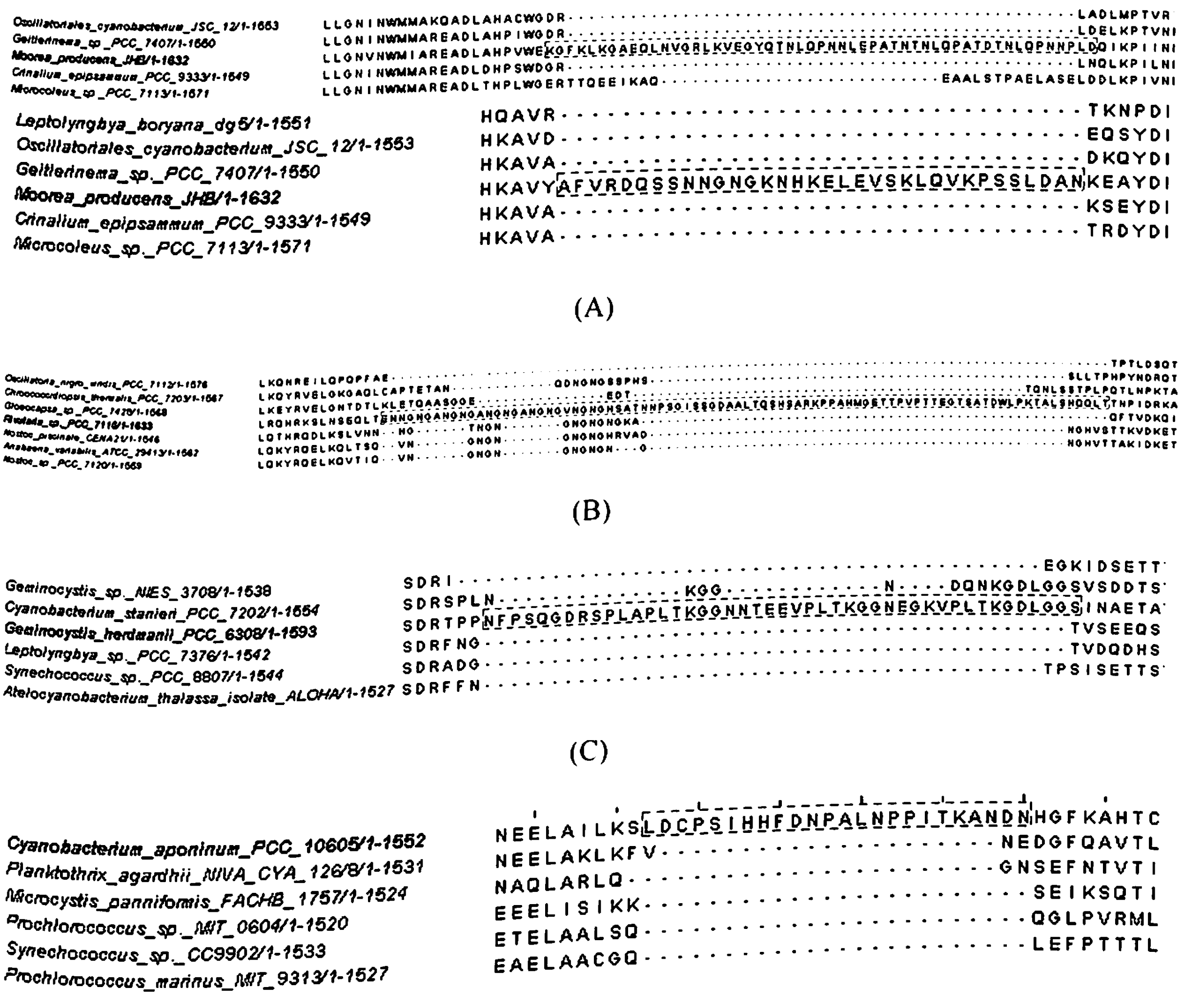
(B)

(C)

(D)

Figure 5.9 Sequence conservation within GOGAT protein. (A) 11 residues within GATase 2 domain (B) 15 residues within Glu_syn_central domain (C) 22 residues within Glu_synthase domain and (D) 11 residue within GXGXG domain.

A multiple sequence alignment (MSA) of 56 protein sequences identified relatively large insertions in GOGAT proteins of *Moorea producens and Rivularia* sp. PCC 711. In *Moorea producens*, two insertions were detected from position 297 to 345 and from 899 to 933, while in *Rivularia* sp. PCC 7116 one insertion was detected from position 468 to 552 (Figure 5.10A and 5.10B respectively). MEME analysis did not show any detectable motif within the inserted region. This protein from several other species (*Geminocystis herdmanii* PCC 6308, Cyanobacterium *aponinum* PCC 10605, *Planktothrix agardhii* NIVA-CYA 126/8, *Cyanobacterium aponinum* PCC 10605 and *Microcystis panniformis* FACHB-1757) also contained several small INDELs (Insertion and Deletion) shown in Figure 5.10C-E. *Arthrospira* species were detected with 15 amino acid long insertions (Figure 5.10F). This insertion in *Arthrospira* appeared to be unique to this genus and could be attributed to the identification of this genus. This insertion likely has some functional role in this protein.

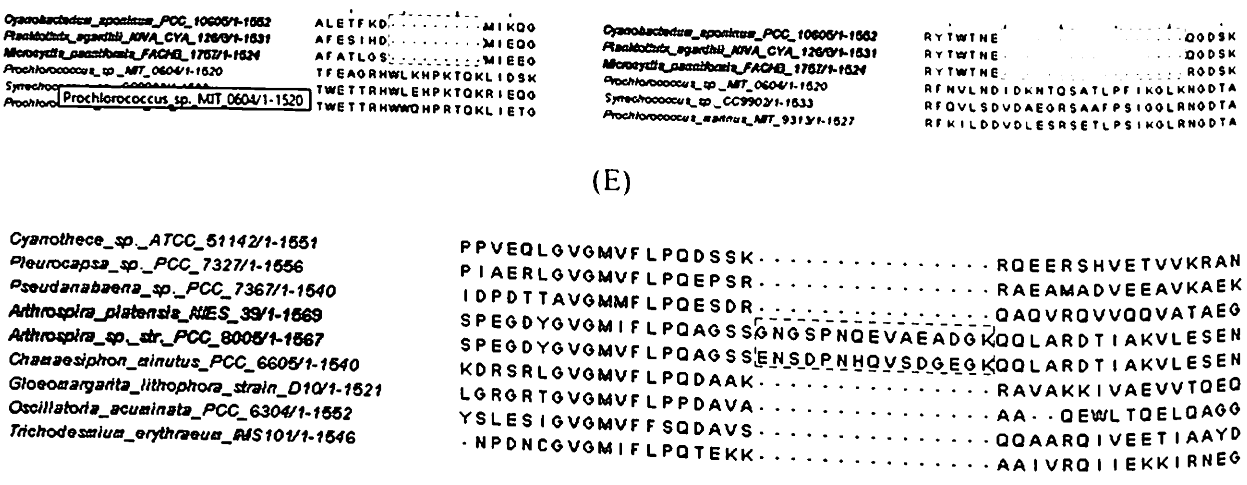

Oscillatoriales_cyanobacterium_JSC_12/1-1553
Geitlerinema_sp._PCC_7407/1-1550
Moorea_producens_JHB/1-1632
Crinalium_epipsammum_PCC_9333/1-1549
Microcoleus_sp._PCC_7113/1-1571
Leptolyngbya_boryana_dg5/1-1551
Oscillatoriales_cyanobacterium_JSC_12/1-1553
Geitlerinema_sp._PCC_7407/1-1550
Moorea_producens_JHB/1-1632
Crinalium_epipsammum_PCC_9333/1-1549
Microcoleus_sp._PCC_7113/1-1571
Geminocystis_sp._NIES_3708/1-1538
Cyanobacterium_stanieri_PCC_7202/1-1554
Geminocystis_herdmanii_PCC_6308/1-1593
Leptolyngbya_sp._PCC_7376/1-1542
Synechococcus_sp._PCC_8807/1-1544
Atelocyanobacterium_thalassa_isolate_ALOHA/1-1527
Cyanobacterium_aponinum_PCC_10605/1-1552
Planktothrix_agardhii_NIVA_CYA_126/8/1-1531
Microcystis_panniformis_FACHB_1757/1-1524
Prochlorococcus_sp._MIT_0604/1-1520
Synechococcus_sp._CC9902/1-1533
Prochlorococcus_marinus_MIT_9313/1-1527

(A)

(B)

(C)

(D)


Cyanobacterium_aponinum_PCC_10605/1-1562
Planktothrix_agardhii_NIVA_CYA_126/8/1-1531
Microcystis_panniformis_FACHB_1757/1-1524
Prochlorococcus_sp._MIT_0604/1-1520
Prochlorococcus sp. MIT_0604/1-1520
ALETFKD··········MIKQQ
AFESIHD··········MIEQQ
AFATLOS··········MIEEO
TFEAORHWLKHPKTQKLIDSK
TWETTRHWLEHPKTQKRIEQQ
TWETTRHWWQHPRTQKLIETQ
Cyanobacterium_aponinum_PCC_10605/1-1562
Planktothrix_agardhii_NIVA_CYA_126/8/1-1531
Microcystis_panniformis_FACHB_1757/1-1524
Prochlorococcus_sp._MIT_0604/1-1520
Synechococcus_sp._CC9902/1-1533
Prochlorococcus_marinus_MIT_9313/1-1527
RYTWTHE··········QQDSK
RYTWTHE··········QQDSK
RYTWTHE··········RQDSK
RFNVLNDIDKNTQSATLPFIKQLKNQDTA
RFQVLSDVDAEQRSAAFPSIQQLRNQDTA
RFKILDDVDLESRSETLPSIKQLRNQDTA
Cyanothece_sp._ATCC_51142/1-1551
Pleurocapsa_sp._PCC_7327/1-1556
Pseudanabaena_sp._PCC_7367/1-1540
Arthrospira_platensis_NIES_39/1-1569
Arthrospira_sp._str._PCC_8005/1-1567
Chamaesiphon_minutus_PCC_6605/1-1540
Gloeomargarita_lithophora_strain_D10/1-1521
Oscillatoria_acuminata_PCC_6304/1-1552
Trichodesmium_erythraeum_IMS101/1-1546
PPVEQLGVGMVFLPQDSSK···············RQEERSHVETVVKRAN
PIAERLGVGMVFLPQEPSR···············RAEAMADVEEAVKAEK
IDPDTTAVGMMFLPQESDR···············QAQVRQVVQQVATAEG
SPEGDYGVGMIFLPQAGSSGNGSPNQEVAEADGKQQLARDTIAKVLESEN
SPEGDYGVGMVFLPQAGSSENSDPNHQVSDGEGKQQLARDTIAKVLESEN
KDRSRLGVGMVFLPQDAAK···············RAVAKKIVAEVVTQEQ
LGRGRTGVGMVFLPPDAVA···············RAVAKKIVAEVVTQEQ
YSLESIGVGMVFFSQDAVS···············AA··QEWLTQELQAGG
·NPDNCGVGMIFLPQTEKK···············QQAARQIVEETIAAYD
AAIVRQIIEKKIRNEG

(E)

(F)

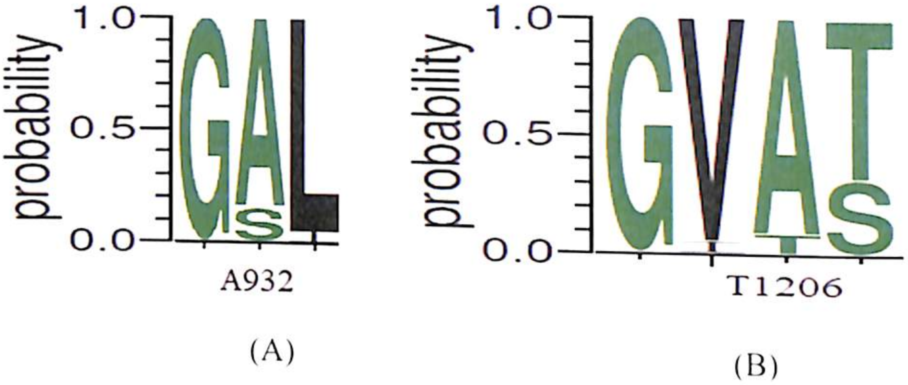
Figure 5.10 Part of the multiple sequence alignment showing the insertions and deletions in different species (A) Two insertions in *Moorea producens* (B) Insertion of *Rivularia* sp. PCC 7116 (C) Insertion of *Geminocystis herdmanii* PCC 6308 (D) Insertion of *Cyanobacterium aponinum* PCC 10605 (E) Two deletions of *Planktothrix agardhii* NIVA-CYA 126/8, *Cyanobacterium aponinum* PCC 10605 and *Microcystis panniformis* FACHB-1757 (F) Insertion of *Arthrospira* species.

Since the crystal structure of GOGAT protein of *Synechocystis* PCC 6803 (PDB ID: 1LLW) was available, we compared all cyanobacterial GOGAT sequences with an available crystal structure (Van den Heuvel et al. 2002). The available crystal structural analysis identified a total of 32 functionally important residues (Table 5.8) that can be divided into three categories viz. (1) FMN binding (19 residues) (2) Iron Sulphur cluster (3Fe-4S) (8 residues) (3) Alpha-ketoglutarate binding (5 residues)

All the above residues were analysed by looking into the MSA and searching for any type of variation within all the selected cyanobacterial species. We found that residues at 2 positions showed variations in which the residue type was different in various species. One residue belonged to Iron Sulphur cluster binding category (1206 of *Arthrospira platensis* NIES-39) while the other one to the alpha-ketoglutarate binding region (932 of *Arthrospira platensis* NIES-39). The sequence conservation at these identified 2 positions is shown in figure 5.11 in the form of sequence logos.

Table 5.8 Variations found in the functionally important residues of GOGAT in cyanobacteria. Functionally important residues were identified by comparing the sequences of *Synechocystis* sp. PCC 6803 and *Arthrospira platensis* NIES-39.

Synechocystis sp. PCC 6803	*Arthrospira platensis* NIES-39	Variations in Cyanobacteria	Type of substitution
FMN binding residues			
G 874	G927	G(48) A(8)	Same
M 875	M928		
S 876	S929		
G 902	G955		
E 903	E956		
Q 944	Q997		
K 966	K1019		
Q 969	Q1022		
K 1034	K1087		
G 1063	G1116		
G 1064	G1117		
T 1065	T1118		
G 1066	G1119		
D 1105	D1158		
G 1106	G1159		
G 1107	G1160	G(53) Q(3)	Differ
G 1128	G1181	G(53) S(3)	Differ
S 1129	T1182	S(49) T(7)	Same
I 1130	I1183	I(40) V(13) A(3)	Same
3Fe-4S (Iron Sulphur cluster binding residues)			
C 1137	C1190		
I 1138	I1191	I(54)Q(1)V(1)	Same
A 1140	A1193	A(53) M(3)	Same
R 1141	R1194		
V 1142	I1195	V(26) I(30)	Same
C 1143	C1196		
C 1148	C1201		
A 1153	T1206	A(51) T(5)	Differ
Alpha-ketoglutarate binding residues			
A 879	A932	A(47) S(9)	Differ
K 972	K1025		
G 977	G1030		
Q 978	Q1031	Q(55) H(1)	Differ
R 992	R1045		



(A)

(B)

Figure 5.11 Sequence variations within functionally important residues at (A) Residue 932 and (B) residue 1206.

5.3.2.2 Phylogenetic analysis

5.3.2.2.1 Gene tree

A gene based NJ tree (Figure 5.12) produced 12 distinct clades. Out of these 12 clades, 5 clades contained species from the same order, i.e. clade 1 contained Nostocales, clades 3 and 5 had Oscillatorials and clade 6 and 8 included species from Synechococales. The remaining 7 clades contained species from different orders. For example, clade 10 had 4 species from 4 orders namely Synechococales, Oscillatorials, Pleurocapsales and Chroccocales. Clade wise comparison of species and gene tree revealed that while most species retained their clades with other co-species in the gene tree as in the species tree, some species moved on to an entirely different clade with different species. We found that three species have changed their positions in the gene tree with respect to the species tree. These were *Planktothrix agardhii* NIVA-CYA 126/8, *Cyanobacterium aponinum* PCC 10605 and *Microcystis panniformis* FACHB-1757. In the species tree, *Planktothrix agardhii* shareed a clade with other species of the Order Oscillatorials with a bootstrap value of 91, *Cyanobacterium aponinum* shared a clade with other species of the Order Chroccocales with a bootstrap value of 100, and the same was observed in the case of *Microcystis panniformis* which shared the clade with *Microcystis aeruginosa* NIES-2549. However, in the gene tree, all the three above mentioned species come closer and formed a single clade (clade 7). These results show that a genetic transfer has occurred between the Order Oscillatorials and Chroccocales which is visible in the gene tree. *Arthrospira platensis* NIES-39 was present in clade 5 of the tree with three other species of the same order which were *Oscillatoria acuminata* PCC 6304, *Oscillatoria nigro-viridis* PCC 7112 and *Trichodesmium erythraeum* IMS101. The remaining species of this order were present in different clades of the tree.

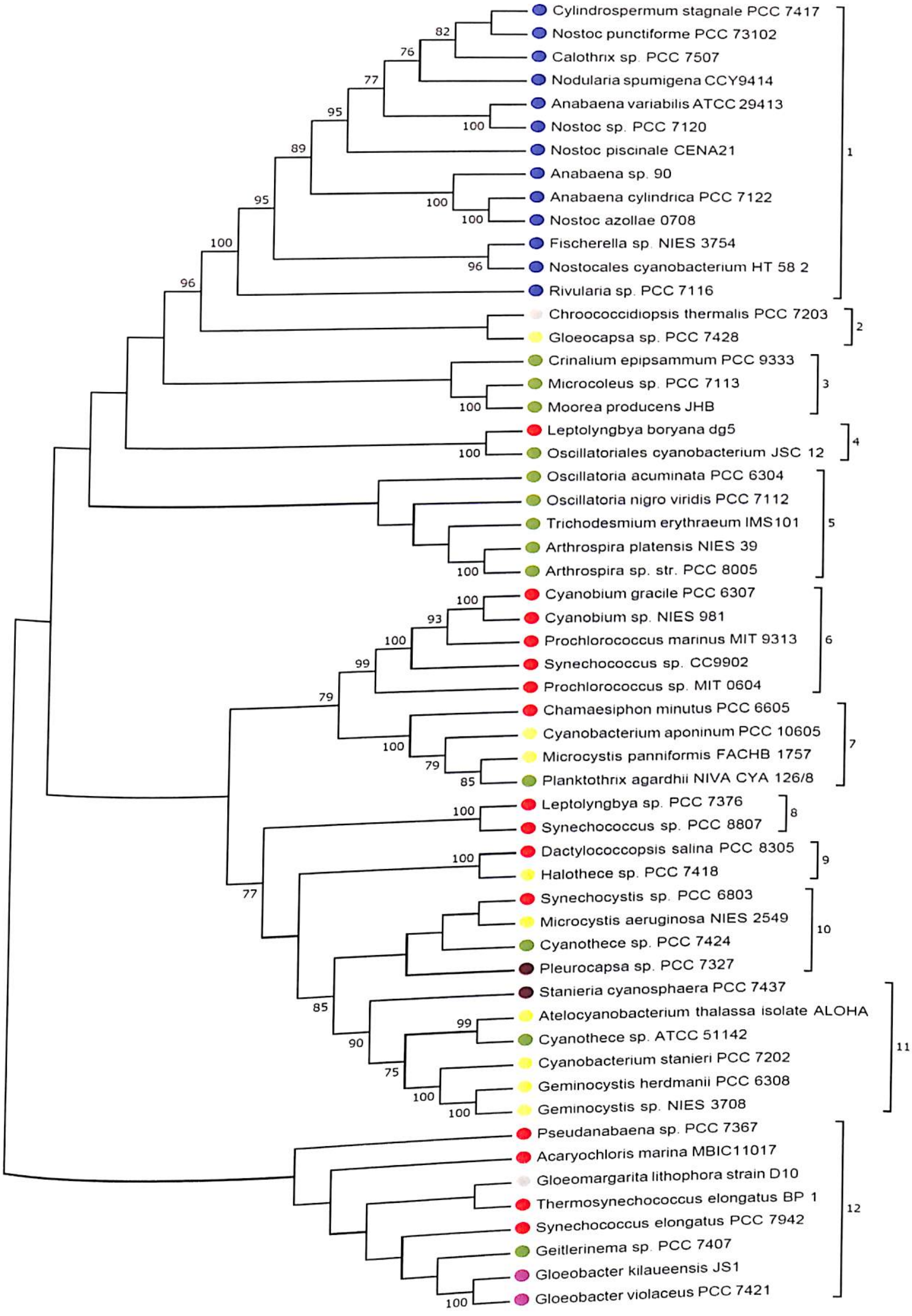

Cylindrospermum stagnale PCC 7417
Nostoc punctiforme PCC 73102
Calothrix sp. PCC 7507
Nodularia spumigena CCY9414
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7120
Nostoc piscinale CENA21
Anabaena sp. 90
Anabaena cylindrica PCC 7122
Nostoc azollae 0708
Fischerella sp. NIES 3754
Nostocales cyanobacterium HT 58 2
Rivularia sp. PCC 7116
Chroococcidiopsis thermalis PCC 7203
Gloeocapsa sp. PCC 7428
Crinalium epipsammum PCC 9333
Microcoleus sp. PCC 7113
Moorea producens JHB
Leptolyngbya boryana dg5
Oscillatoriales cyanobacterium JSC 12
Oscillatoria acuminata PCC 6304
Oscillatoria nigro viridis PCC 7112
Trichodesmium erythraeum IMS101
Arthrospira platensis NIES 39
Arthrospira sp. str. PCC 8005
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
Prochlorococcus marinus MIT 9313
Synechococcus sp. CC9902
Prochlorococcus sp. MIT 0604
Chamaesiphon minutus PCC 6605
Cyanobacterium aponinum PCC 10605
Microcystis panniformis FACHB 1757
Planktothrix agardhii NIVA CYA 126/8
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 8807
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Synechocystis sp. PCC 6803
Microcystis aeruginosa NIES 2549
Cyanothece sp. PCC 7424
Pleurocapsa sp. PCC 7327
Stanieria cyanosphaera PCC 7437
Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Cyanobacterium stanieri PCC 7202
Geminocystis herdmanii PCC 6308
Geminocystis sp. NIES 3708
Pseudanabaena sp. PCC 7367
Acaryochloris marina MBIC11017
Gloeomargarita lithophora strain D10
Thermosynechococcus elongatus BP 1
Synechococcus elongatus PCC 7942
Geitlerinema sp. PCC 7407
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
82
76
77
95
100
89
100
100
95
100
96
96
100
100
100
100
100
93
100
99
79
100
79
85
100
100
77
99
85
90
75
100
100
100
1
2
3
4
5
6
7
8
9
10
11
12

Figure 5.12 GOGAT gene-based NJ tree of 56 cyanobacterial species contains 12 distinct clades. Color coding is same as figure 4.1.

5.3.2.2.2 Protein tree

Protein tree (Figure 5.13) is much more conserved than the gene tree. Among the 14 distinct clades, 8 clades contain species from the same order, i.e. clade 1 had Nostocales, clades 4 and 5 included Oscillatorials, clade 7 contained Chroccocales, clade 8, 13 and 14 included Synechococales and clade 11 of Gloeobacterales. The remaining 6 clades contained species from different orders. For example, clade 9 had 6 species from 4 different orders namely Oscillatorials, Chroccocales, Synechococales and Pleurocapsales. It was observed that in the species tree, *Planktothrix agardhii* shared a clade with other species of the Order Oscillatorials with a bootstrap value of 91, *Cyanobacterium aponinum* shared a clade with other species of the Order Chroccocales with a bootstrap value of 100, and the same was observed in the case of *Microcystis panniformis* which shared the clade with *Microcystis aeruginosa* NIES-2549 with a bootstrap value of 100. However, in the protein tree, all three species came closer and formed a single clade (clade 12). This clade had a high bootstrap of 99 and 100 again. These results also confirmed a horizontal gene transfer event between the Order Oscillatorials and Chroccocales. *Arthrospira platensis* NIES-39 was present in clade 5 with the same species as was observed in the gene tree.

5.3.2.3 Codon usages

We compared gene and protein trees to find any evidence of different codon usages. These two trees were very similar regarding the topology and the placement of species in various clades. Interestingly, not a single species shared clade with different species in the gene tree and protein tree. We also made a gene tree based on the first 2 codons. This tree also showed similar topology as that of the gene and the protein tree. This similarity reconfirms the highly conserved nature of GOGAT protein among cyanobacterial species.

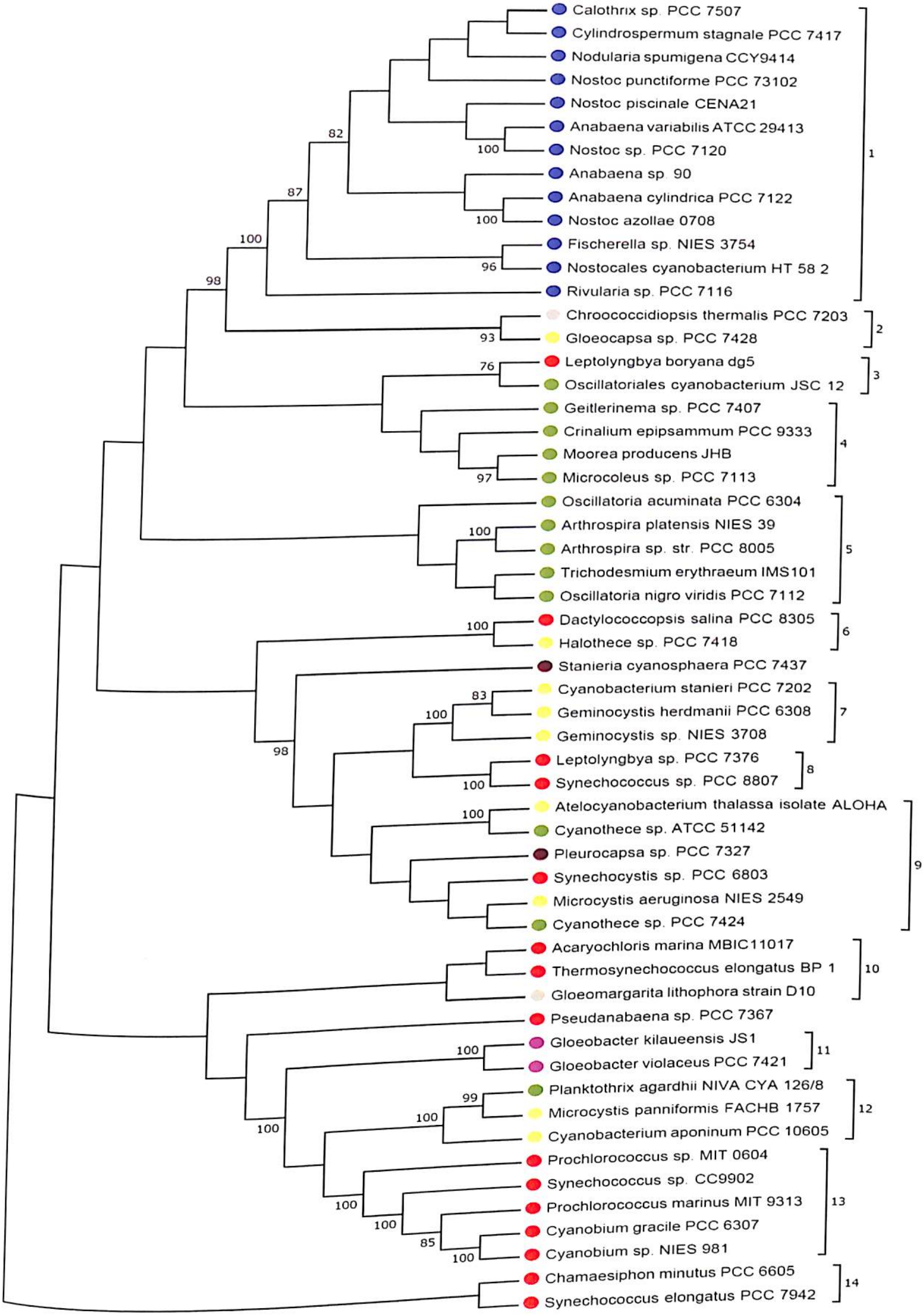

Calothrix sp. PCC 7507
Cylindrospermum stagnale PCC 7417
Nodularia spumigena CCY9414
Nostoc punctiforme PCC 73102
Nostoc piscinale CENA21
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7120
Anabaena sp. 90
Anabaena cylindrica PCC 7122
Nostoc azollae 0708
Fischerella sp. NIES 3754
Nostocales cyanobacterium HT 58 2
Rivularia sp. PCC 7116
Chroococcidiopsis thermalis PCC 7203
Gloeocapsa sp. PCC 7428
Leptolyngbya boryana dg5
Oscillatoriales cyanobacterium JSC 12
Geitlerinema sp. PCC 7407
Crinalium epipsammum PCC 9333
Moorea producens JHB
Microcoleus sp. PCC 7113
Oscillatoria acuminata PCC 6304
Arthrospira platensis NIES 39
Arthrospira sp. str. PCC 8005
Trichodesmium erythraeum IMS101
Oscillatoria nigro viridis PCC 7112
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Stanieria cyanosphaera PCC 7437
Cyanobacterium stanieri PCC 7202
Geminocystis herdmanii PCC 6308
Geminocystis sp. NIES 3708
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 8807
Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Pleurocapsa sp. PCC 7327
Synechocystis sp. PCC 6803
Microcystis aeruginosa NIES 2549
Cyanothece sp. PCC 7424
Acaryochloris marina MBIC11017
Thermosynechococcus elongatus BP 1
Gloeomargarita lithophora strain D10
Pseudanabaena sp. PCC 7367
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Planktothrix agardhii NIVA CYA 126/8
Microcystis panniformis FACHB 1757
Cyanobacterium aponinum PCC 10605
Prochlorococcus sp. MIT 0604
Synechococcus sp. CC9902
Prochlorococcus marinus MIT 9313
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
Chamaesiphon minutus PCC 6605
Synechococcus elongatus PCC 7942
82
100
87
100
100
96
98
93
76
97
100
100
83
100
98
100
100
100
99
100
100
100
100
85
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14

Figure 5.13 GOGAT protein tree with 14 distinct clades. Colour coding is the same as figure 4.1.

5.3.2.4 Gene Duplication and Speciation events

Even in the case of GOGAT we found extensive gene duplication and speciation events supported by a good bootstrap value (≥75%) (Figure 5.14). For example, *Leptolyngbya boryana* dg5 (extremophile) and *Oscillatoriales cyanobacterium* JSC 12 (normal fresh water) were present close to each other (bootstrap = 100%) despite belonging to different orders, hence proving their common origin. Similarly, *Dactylococcopsis salina* PCC 8305 and *Halothece* sp. PCC 7418, *Cyanothece* sp. ATCC 51142 (photosynthetic) and *Atelocyanobacterium thalassa* isolate ALOHA (non-photosynthetic) and also *Planktothrix agardhii* NIVA-CYA 126/8 (fresh water), *Cyanobacterium aponinum* PCC 10605 (thermal springs) and *Microcystis panniformis* FACHB-1757 (freshwater) showed the same behaviour in this protein. These observations support the widespread diversity of cyanobacterial species and the effect of evolutionary pressure on the evolution of this protein. However, comparing the speciation events of both GS and GOGAT, we observed that similar species are involved in the speciation event in both the proteins, which indicates that these two proteins do not contribute much in the speciation event.

5.3.2.5 Structural analysis

To look into the 3-Dimensional structure of two identified residues which showed variation among cyanobacteria and to analyse the structure of the insertion identified in *Arthrospira* genus, we modeled the representative species of the clades obtained in the GOGAT protein tree using the Modeler v9.15 (Table 5.9). We modeled nine species which belonged to the four major Orders of the cyanobacteria, which cover 50 out of 56 total species selected. The template used was the crystal structure of Glutamate Synthase from *Synechocystis* sp. PCC 6803 (PDB-1LLW) with a query coverage ranging from 92 to 98% and identity between 45 and 73%. The best model (based on N-DOPE score) was energy minimised and was validated using the Verify3D, ERRAT, Qmean score and WhatCheck programs. The results of these validations are shown in Table 5.10.

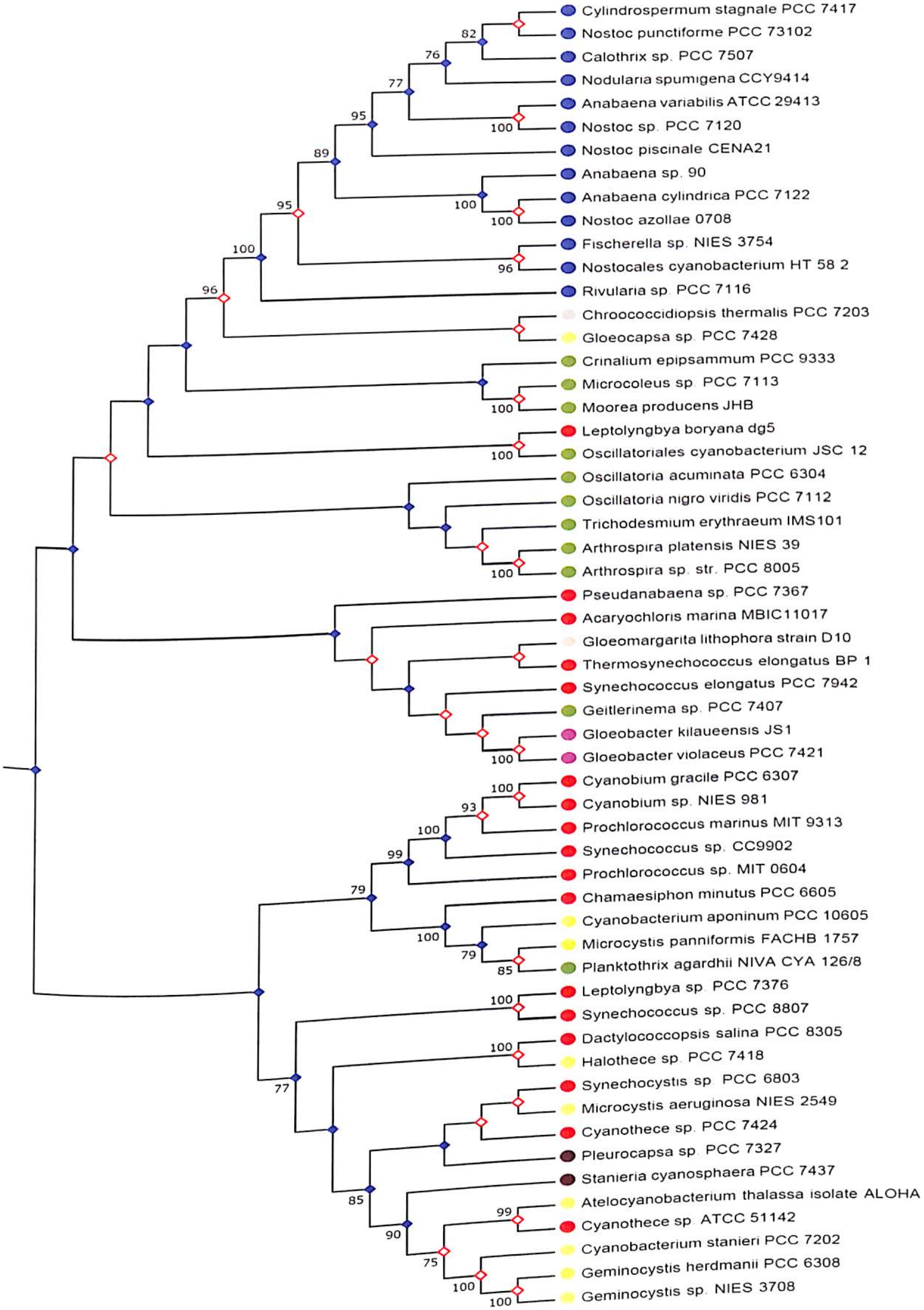

Cylindrospermum stagnale PCC 7417
Nostoc punctiforme PCC 73102
Calothrix sp. PCC 7507
Nodularia spumigena CCY9414
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7120
Nostoc piscinale CENA21
Anabaena sp. 90
Anabaena cylindrica PCC 7122
Nostoc azollae 0708
Fischerella sp. NIES 3754
Nostocales cyanobacterium HT 58 2
Rivularia sp. PCC 7116
Chroococcidiopsis thermalis PCC 7203
Gloeocapsa sp. PCC 7428
Crinalium epipsammum PCC 9333
Microcoleus sp. PCC 7113
Moorea producens JHB
Leptolyngbya boryana dg5
Oscillatoriales cyanobacterium JSC 12
Oscillatoria acuminata PCC 6304
Oscillatoria nigro viridis PCC 7112
Trichodesmium erythraeum IMS101
Arthrospira platensis NIES 39
Arthrospira sp. str. PCC 8005
Pseudanabaena sp. PCC 7367
Acaryochloris marina MBIC11017
Gloeomargarita lithophora strain D10
Thermosynechococcus elongatus BP 1
Synechococcus elongatus PCC 7942
Geitlerinema sp. PCC 7407
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Cyanobium gracile PCC 6307
Cyanobium sp. NIES 981
Prochlorococcus marinus MIT 9313
Synechococcus sp. CC9902
Prochlorococcus sp. MIT 0604
Chamaesiphon minutus PCC 6605
Cyanobacterium aponinum PCC 10605
Microcystis panniformis FACHB 1757
Planktothrix agardhii NIVA CYA 126/8
Leptolyngbya sp. PCC 7376
Synechococcus sp. PCC 8807
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Synechocystis sp. PCC 6803
Microcystis aeruginosa NIES 2549
Cyanothece sp. PCC 7424
Pleurocapsa sp. PCC 7327
Stanieria cyanosphaera PCC 7437
Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp. ATCC 51142
Cyanobacterium stanieri PCC 7202
Geminocystis herdmanii PCC 6308
Geminocystis sp. NIES 3708
82
76
77
95
100
89
100
95
100
100
96
96
100
100
100
100
100
93
100
99
79
100
79
85
100
100
77
85
99
90
75
100
100

Figure 5.14 Evolutionary relationships among taxa shows 15 significant (with bootstrap value >75%) gene duplication events (closed diamonds) and 18 significant speciation events (open diamonds).

Table 5.9 Modeled species of the representative clades of GOGAT protein tree. Template used was GOGAT of *Synechocystis* sp. PCC 6803 (PDB-1LLW).

Species modelled	Protein length	Query Coverage (%)	Identity (%)
Arthrospira platensis NIES-39	1569	96	66
Dactylococcopsis salina PCC 8305	1547	97	67
Geminocystis sp. NIES-3708	1538	97	71
Gloeocapsa sp. PCC 7428	1568	97	70
Anabaena variabilis ATCC 29413	1562	97	68
Planktothrix agardhii NIVA-CYA 126/8	1531	97	45
Rivularia sp. PCC 7116	1633	92	70
Cyanothece sp. PCC 7424	1551	98	73
Cyanobium gracile PCC 6307	1534	97	61

Table 5.10 The quality of the predicted GOGAT structures was estimated through various servers which were considered as good structures.

Species	Verify3D	Errat	Q-mean	WhatCheck
Arthrospira platensis NIES-39	88.67	87.30	-2.79	Pass
Dactylococcopsis salina PCC 8305	89.63	87.26	-2.38	Pass
Geminocystis sp. NIES-3708	85.76	84.91	-2.04	Pass
Gloeocapsa sp. PCC 7428	87.59	85.09	-2.37	Pass
Anabaena variabilis ATCC 29413	88.87	86.54	-2.06	Pass
Planktothrix agardhii NIVA-CYA 126/8	88.34	83.96	-3.44	Pass
Rivularia sp. PCC 7116	90.27	85.99	-2.35	Pass
Cyanothece sp. PCC 7424	89.87	87.41	-2.21	Pass
Cyanobium gracile PCC 6307	88.95	89.29	-3.06	Pass

After superimposition of all the modelled structures, it was observed that the majority of the functionally important residues were conserved in the cyanobacteria in terms of orientation in the 3-dimensional space. However as mentioned above, there are two positions at which the variations in terms of amino acid composition was found in various cyanobacteria. These positions are A932 which is involved in alpha-ketoglutarate binding and T1206 which is required for Iron Sulphur cluster (3Fe-4S) binding (numbered according to *Arthrospira* *platensis* NIES-39). The variation among these two positions includes Alanine (47) and Serine (9) at position 932 and Alanine (51) and Threonine (5) at position 1206 respectively (in the bracket – the number of species having that amino acid out of total 56 species). In an attempt to identify specific features of GOGAT of *Arthrospira platensis* NIES-39 that confer it the ability to produce high protein content, we observed that Threonine at position 1206 was present in only 5 species out of a total of 56 species. These species were *Arthrospira platensis* NIES-39, *Arthrospira* sp. str. PCC 8005, *Geitlerinema* sp. PCC 7407, *Pseudanabaena* sp. PCC 7367 and *Oscillatoria nigro-viridis* PCC 7112. This observation hinted that the Order Oscillatorials was diverse since four of the above species belonged to the Order Oscillatorials. The variation at this position is highlighted in figure 5.15 as a superimposition of modeled structures of *Arthrospira platensis* NIES-39 and *Synechocystis* PCC 6803. The structural analysis revealed that the 15 amino acid long insertion in *Arthrospira platensis* (Figure 5.16A) was present in the GATase2 domain. The *ab-initio* method-based Quark tool (Xu and Zhang 2012) predicted a possible single helical structure for the inserted region (Figure 5.16B). As the GATase2 domain is involved in the binding of Glutamine, this insertion could play an important role in the GOGAT protein function of *Arthrospira platensis* NIES-39.

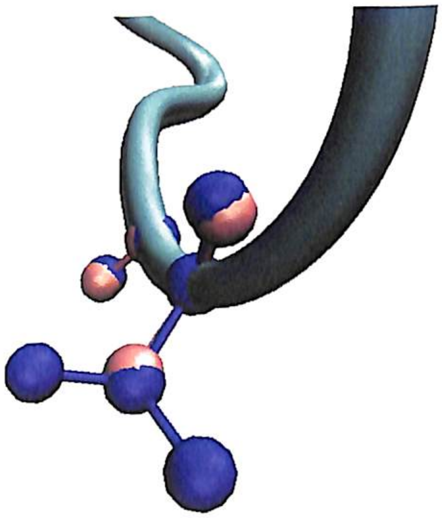


Figure 5.15 Superimposed structure of *Synechocystic* PCC 6803 (PDB-1LLW) and modeled *Arthrospira platensis* NIES-39 showing the variation among the two cyanobacterial species at position 1206 of *Arthrospira platensis* NIES-39 with Alanine (pink) and Threonine (blue).

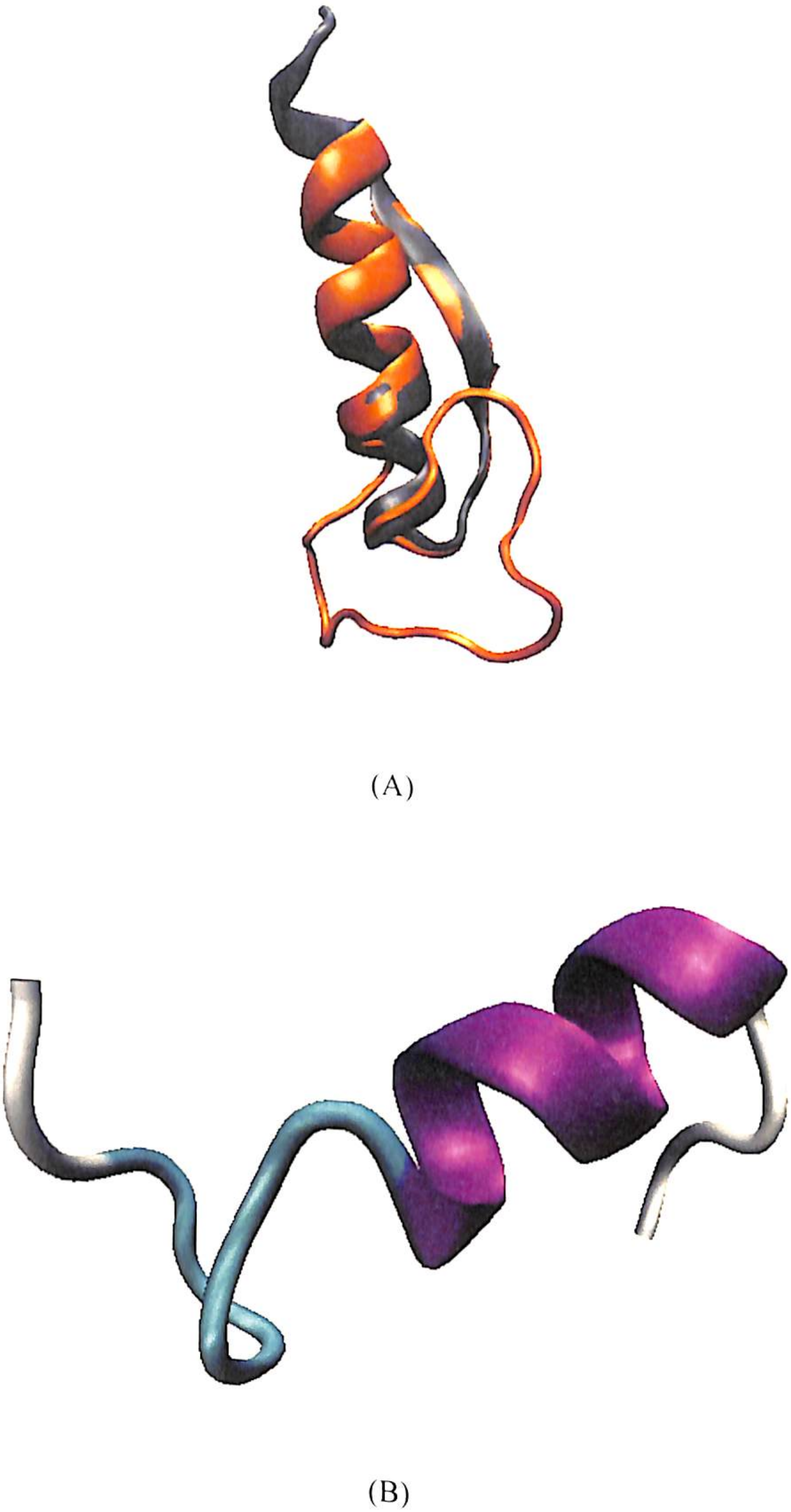


Figure 5.16 (A) An insertion of 15 residues is identified within the GATase 2 domain of *Arthrospira platensis* (B) The predicted structure of inserted region is shown.

5.4 Conclusions

This study compares the GS/GOGAT pathway enzymes of *Arthrospira platensis* NIES-39 with other cyanobacteria in terms of sequence, structure and evolution. This pathway has two enzymes, i.e. glutamine synthetase and glutamate synthase and helps in the incorporation of nitrogen in various biologically important biomolecules.

For Glutamine synthetase, we identified the signature pattern of the domains present in this enzyme within cyanobacteria. Functionally important residues were identified in the GS of *Arthrospira platensis* NIES-39. This enzyme was highly conserved and showed very little sequence or structural variations with respect to GS from other cyanobacteria. Phylogenetic analysis also revealed the conserved nature of this enzyme. Codon usages were identified in some species of GS. Significant speciation events were identified in this enzyme.

Signature patterns were also identified for the domains of glutamate synthase. Our sequence analysis had identified a 15 amino acids long insertion in *Arthrospira* species. This insertion in *Arthrospira* is present in the GATase 2 domain and is unique to this genus. An α-helix has been predicted in this region and could be assigned a functional role in this protein. Phylogenetic analysis revealed that GOGAT have a different evolutionary pattern in some species like *Planktothrix agardhii* NIVA-CYA 126/8, *Cyanobacterium aponinum* PCC 10605 and *Microcystis panniformis* FACHB-1757. We have also identified deletions in these three species which could be related to the closeness of these species. Analysis of functionally important residues identified a key residue Threonine 1206 in *Arthrospira platensis* NIES-39. The corresponding position in *Synechocystis* PCC 6803, a cyanobacterium with low protein content is occupied by Alanine. These residues have different natures and also showed different orientations in their 3-dimentional structures which could result in differential functioning of this enzyme in different species.

Chapter VI

Conclusion and Future perspectives

6.1 Conclusion

Arthrospira platensis NIES-39 is a non-nitrogen fixing and filamentous cyanobacterium. It is a photosynthetic prokaryote with high photosynthetic efficiency and hence it contributes significantly to the nitrogen and carbon cycle. It is also used as a food supplement due to its high nutritional values, particularity because of its high protein content. Apart from using as food, it is also used in many other fields of science like nanobiotechnology, biosensors, biofuel and biofertilizers which makes it a commercially important species. It is also an alkalophilic and halophilic organism. Due to all its features it is considered to be a connecting link between prokaryotes and eukaryotes and hence makes an ideal system for study.

Hence, taking *Arthrospira platensis* NIES-39 as our species of study, we worked upon two objectives, i.e. functional annotation of the hypothetical proteins of *Arthrospira platensis* NIES-39 genome and secondly, to look into the role of nitrogen assimilation pathway enzymes in the high protein content of this cyanobacterium.

Annotating a protein is crucial as it gives us information about various parameters like the function, structure, location, its physical parameters, possible mechanisms of action, pathway involvements and interactions with other molecules. *Arthrospira platensis* NIES-39, being an alkalophilic, halophilic and high protein containing species, it is of immense significance to annotate the hypothetical proteins as these proteins may give a clue regarding molecular basis of observed characteristics features of this organism.

In the present study, hypothetical proteins of *Arthrospira platensis* NIES-39 genome were annotated using *in silico* approaches. We have used a defined pipeline of various computational methods to annotate the proteins with their respective functions. With our defined method we were able to functionally annotate 526 proteins out of a total of 1364. This analysis was able to annotate a variety of different functionally important proteins like DNA binding proteins, endonucleases, ATP binding proteins, transcriptional regulators and proteins involved in carbohydrate metabolism which play a vital role in different cellular mechanisms like gene expression and regulation, DNA repair, energy generation and metabolite transport. A substantial proportion of the annotated protein (10%) we identified as membrane proteins. These membrane proteins play a crucial role in the adaptation of the organisms to the adverse environments, particularly in high salt and high pH conditions. Some stress related proteins have also been identified like alpha crystalline family protein and nirD stress tolerance protein. Annotation of these proteins can definitely help our understanding about the mechanism behind the alkalophilic and halophilic nature of

Arthrospira platensis NIES-39. In addition, we have also annotated few proteins that are related to the translational machinery like amino acid-tRNA ligase activity, peptidase activity, amino acids metabolism and a nitrate reductase associated protein. These proteins could play a functional role in determining the protein content of a cell as in *Arthrospira platensis* NIES-39. Protein-protein interaction is an important aspect of cell functioning as most of the cellular proteins interact with each other. These interactions can tell us about the probable pathway a protein may involved in. Our annotated proteins highly interacted among each other indicating their role in stress tolerance as majority of the proteins belonged to membrane proteins and enzymatic activity. Finally, we can say that the functional annotation of hypothetical proteins of *Arthrospira platensis* NIES-39 may help in understanding the various potential stress induced proteins and in understanding the mechanism behind the high protein content.

Nitrogen is a crucial element in a cell's life. Vital biomolecules like nucleic acids and proteins both contain nitrogen. Biosynthetic pathways of these biomolecules receive nitrogen from the nitrogen assimilatory pathway. Four enzymes (NR, NiR, GS and GOGAT) are present in this pathway. This pathway can be broadly categorized into two sub-pathways i.e. nitrate assimilation where the absorbed nitrate gets converted into ammonium via nitrate reductase (NR) and nitrite reductase (NiR) and the second is the well-known GS-GOGAT pathway through which the ammonium formed gets incorporate into various biomolecules via glutamine and glutamate.

In this study, we have looked into the sequence and structural features of these enzymes in *Arthrospira platensis* NIES-39 and searched for any variations that have significant impact on the nitrogen assimilation process. Thus, having different functionality, we could relate it to the protein content of *Arthrospira platensis* NIES-39.

Signature sequences are a unique pattern of amino acid residues which are characteristic features of a group of sequences. In this study, we have identified the signature pattern of the domains of both the nitrate assimilatory enzymes i.e. nitrate reductase and nitrite reductase within cyanobacteria. These signature patterns can uniquely identify the cyanobacterial class and could be helpful in identifying new homologs of this protein in other cyanobacteria. Motifs represent secondary structure that helps the protein in proper functioning. They could alter the protein function and make it more efficient. In our analysis, the enzyme nitrate reductase was detected with a possible motif with α-helical geometry at the C-terminal of the protein. This motif could enhance this enzyme's stability and its contribution towards the final protein content.

Active site of a protein contains functionally important residues. These residues tend to remain conserved in the homologous sequences. Any change/mutation in these residues will affect the protein functioning. We analyzed these functionally important residues in NR and NiR within cyanobacteria. Our analysis was able to uniquely identify the key residues in both the enzymes which would affect the protein functioning. In case of NR of *Arthrospira platensis* NIES-39, the position 394 is involved in guiding the nitrate towards the active site, was mutated from Serine in *Synechocystis* sp. PCC 6803 to Asparagine in *Arthrospira platensis* NIES-39. This replacement could enhance the capability of the enzyme to get more substrate and hence more product. A similar kind of phenomenon was also observed in NiR, where the active site residue position 408 has changed from Lysine in *Synechocystis* sp. PCC 6803 to Asparagine in *Arthrospira platensis* NIES-39. This position was already known to switch this protein from high to low affinity in Tobacco. This is the first report of the dual nature of NiR in cyanobacteria. Asparagine in *Arthrospira platensis* NIES-39 makes NiR a low affinity enzyme increasing its turn over number.

Tertiary structures immensely help us in understanding the working of a protein. We used homology modeling to model the representative species of each order from the protein trees of both NR and NiR. This study helps us to look into the 3-dimentional structure of the functionally important residues, particularly the ones which showed variations. From this analysis, we can tell that the key residues identified in NR and NiR have different orientations and could functionally affect the enzymes.

The evolutionary pattern of these enzymes was also studied through phylogenetic analysis. Species, gene and protein tree were constructed. 16s rRNA gene-based species tree revealed that there is a gap between the classical and modern approaches of taxonomy. Comparing species and gene tree predicted high speciation events among all cyanobacteria which support their wide geographical presence. Gene and protein tree comparison gives us the idea about the codon usages. Here also we found codon usages in some of the species like *Gloeocapsa* sp. PCC 7428 and *Oscillatoriales cyanobacterium* in NR and *Cyanobium* and *Prochlorococcus* in NiR. Some species shows different evolutionary pattern when species and protein tree was compared in both NR and NiR. However, *Arthrospira platensis* NIES-39 showed conserved evolutionary pattern.

Sequence analysis of the GS and GOGAT in terms of signature patterns revealed that GS is a highly conserved enzyme as compared to GOGAT. Motif analysis also hints at the same conclusion with only 8 highly conserved motifs identified in GS as compared to 20 motifs in GOGAT. Comparing the homologous sequences of cyanobacteria, an insertion was identified in the GATase domain of GOGAT. This insertion has an α-helix. GATase domain is involved in the Glutamine binding and presence of this insertion could affect the enzyme functioning. Functionally important residues were identified and analyzed in both GS and GOGAT. As expected, no variation was detected in highly conserved GS. While in case of *Arthrospira platensis* NIES-39 GOGAT, a key functional residue at position 1206 was identified which is involved in Iron-Sulphur cluster binding. The change from Alanine to Threonine at this position could affect the enzymatic activity of GOGAT.

The modeled species from representative orders of GS and GOGAT confirmed the conserved nature of GS. GS of 8 modeled species were superimposed and the structures were highly concurrent. In case of GOGAT, the amino acids at position 1206 were different in terms of their orientations and likely to affect the protein function.

Phylogenetic analysis revealed a horizontal gene transfer event between the Order Oscillatorials and Chroccocales. High speciation events were also detected. Here also we found codon usages in some of the species in GS, but codon usages were not found in GOGAT.

Thus, the present study gives us an idea on the various proteins that may play an important role in stress tolerance and protein content of *Arthrospira platensis* NIES-39. It also gives a possible idea about the sequence and structural features of the enzymes of nitrogen assimilation pathway of *Arthrospira platensis* NIES-39 that could affect the enzyme function and eventually can lead to high protein content. However, the study of in-depth molecular events functioning in this process is still at the research level.

6.2 Future Perspectives

In this study, we have annotated 312 un-annotated proteins of *Arthropsira platensis* NIES-39. Key proteins involved in stress management have been identified. These proteins can further be investigated for their individual contribution in the stress tolerance. Experimental studies like proteomics analysis can really help us to understand the various mechanisms in stress tolerance.

Proteins involved in translational process have also been identified. Experiments can be set up to find the molecular mechanism behind the unique characteristic features of *Arthropsira platensis* NIES-39 like the high protein content.

This annotation process can be used in the annotation process of other newly sequenced genomes using *in silico* approaches like holology searching.

With the annotation of new interacting proteins, metabolic pathway analysis using methods like Flux Balance Analysis could be helpful in identifying novel pathways contibuting towards protein content. These putative pathways can be validated using experiments.

Nitrite reductase of *Arthropsira platensis* NIES-39 was identified as a dual-affinity enzyme. More studies like molecular dynamics and simulations can help to find the actual mechanism behind this process.

Several pathways have been known to affect the final protein content of a cell. These pathways include mRNA degradation and tRNA synthetase. Looking deeply into the working of these pathways will give us new perspectives to think about the high protein content of the cell.

Although we have tried to take as many species as possible to include variations, new species have been sequenced every day and hence including more species would definitely enhance the variation among dataset and thus chances of getting new insights into the current process will be more.

References

Abdulqader, G., L. Barsanti and M. R. Tredici (2000). "Harvest of Arthrospira platensis from Lake Kossorom (Chad) and its household usage among the Kanembu." Journal of Applied Phycolology 12: 493–498.

Abed, R. M., S. Dobretsov and K. Sudesh (2009). "Applications of cyanobacteria in biotechnology." J Appl Microbiol 106(1): 1-12.

Ajayan, K. V. (2011). "Response of Temperature and pH on the Growth and Biochemical changes in Spirulina platensis " J. Biol. – Plant Biol. 56(1): 37-42.

Ali, A., P. Jha, K. S. Sandhu and N. Raghuram (2008). "Spirulina nitrate-assimilating enzymes (NR, NiR, GS) have higher specific activities and are more stable than those of rice." Physiol Mol Biol Plants 14(3): 179-182.

Allakhverdiev, S. I., M. Kinoshita, M. Inaba, I. Suzuki and N. Murata (2001). "Unsaturated fatty acids in membrane lipids protect the photosynthetic machinery against salt-induced damage in Synechococcus." Plant Physiol 125(4): 1842-1853.

Allen, A. E., M. G. Booth, M. E. Frischer, P. G. Verity, J. P. Zehr and S. Zani (2001). "Diversity and detection of nitrate assimilation genes in marine bacteria." Appl Environ Microbiol 67(11): 5343-5348.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol 215(3): 403-410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res 25(17): 3389-3402.

Apse, M. P. and E. Blumwald (2002). "Engineering salt tolerance in plants." Curr Opin Biotechnol 13(2): 146-150.

Bai, N. J. (1985). "Competitive exclusion or morphological transformation? A case study with Spirulina fusiformis." Algological Studies/Archiv für Hydrobiologie Supplement Volumes: 191-199.

Bai, N. J. and C. V. Seshadri (1983). "On Coiling and Uncoling of Trichomes Iin the Genus Spirulina." SCHWEIZERISCHE ZEITSCHRIFT FUR HYDROLOGIE-SWISS JOURNAL OF HYDROLOGY 45(1): 297-298.

Bailey, T. L., N. Williams, C. Misleh and W. W. Li (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." Nucleic Acids Res 34(Web Server issue): W369-373.

Baldwin, S. A. (1993). "Mammalian passive glucose transporters: members of an ubiquitous family of active and passive transport proteins." Biochim Biophys Acta 1154(1): 17-49.

Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats and S. R. Eddy (2004). "The Pfam protein families database." Nucleic Acids Res 32(Database issue): D138-141.

Baylan, M., B. D. Özcan, O. ISIK, M. AKAR and S. Yazar (2012). "A Mini Review on Spirulina." Turkish Journal of Scientific Reviews 1: 031-034.

Belasco, J. G., G. Nilsson, A. Von Gabain and S. N. Cohen (1986). "The stability of E. coli gene transcripts is dependent on determinants localized to specific mRNA segments." Cell 46(2): 245-251.

Benkert, P., M. Biasini and T. Schwede (2011). "Toward the estimation of the absolute quality of individual protein structure models." Bioinformatics 27(3): 343-350.

Benkert, P., M. Kunzli and T. Schwede (2009). "QMEAN server for protein model quality estimation." Nucleic Acids Res 37(Web Server issue): W510-514.

Benkert, P., S. C. Tosatto and D. Schomburg (2008). "QMEAN: A comprehensive scoring function for model quality assessment." Proteins 71(1): 261-277.

Berg, J. M., J. L. Tymoczko, L. Stryer and G. J. J. Gatto (2012). Biochemistry. New York, W.H. Freeman & Company.

Bergman, B., G. Sandh, S. Lin, J. Larsson and E. J. Carpenter (2013). "Trichodesmium--a widespread marine cyanobacterium with unusual nitrogen fixation properties." FEMS microbiology reviews 37(3): 286-302.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000). "The Protein Data Bank." Nucleic Acids Res 28(1): 235-242.

Blaha, L., P. Babica and B. Marsalek (2009). "Toxins produced in cyanobacterial water blooms - toxicity and risks." Interdiscip Toxicol 2(2): 36-41.

Blumwald, E., G. S. Aharon and M. P. Apse (2000). "Sodium transport in plant cells." Biochim Biophys Acta 1465(1-2): 140-151.

Brenchley, R., M. Spannagl, M. Pfeifer, G. L. Barker, R. D'Amore, A. M. Allen, N. McKenzie, M. Kramer, A. Kerhornou, D. Bolser, S. Kay, D. Waite, M. Trick, I. Bancroft, Y. Gu, N. Huo, M. C. Luo, S. Sehgal, B. Gill, S. Kianian, O. Anderson, P. Kersey, J. Dvorak, W. R. McCombie, A. Hall, K. F. Mayer, K. J. Edwards, M. W. Bevan and N. Hall (2012). "Analysis of the bread wheat genome using whole-genome shotgun sequencing." Nature 491(7426): 705-710.

Brown, J. R., Y. Masuchi, F. T. Robb and W. F. Doolittle (1994). "Evolutionary relationships of bacterial and archaeal glutamine synthetase genes." J Mol Evol 38(6): 566-576.

Burillo, S., I. Luque, I. Fuentes and A. Contreras (2004). "Interactions between the nitrogen signal transduction protein PII and N-acetyl glutamate kinase in organisms that perform oxygenic photosynthesis." Journal of Bacteriology 186(11): 3346-3354.

Burton, Z. F., G. A. Carol, K. K. Watanabe and R. R. Burgess (1983). "The operon that encodes the sigma subunit of RNA polymerase also encodes ribosomal protein S21 and DNA primase in E. coli K12." Cell 32(2): 335-349.

Campbell, W. H. (1999). "NITRATE REDUCTASE STRUCTURE, FUNCTION AND REGULATION: Bridging the Gap between Biochemistry and Physiology." Annu Rev Plant Physiol Plant Mol Biol 50: 277-303.

Castenholz, R. W. (2001). The Archaea and the Deeply Branching, and Phototrophic Bacteria New York, Springer.

Cavalier-Smith, T. (2002). "The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification." Int J Syst Evol Microbiol 52(Pt 1): 7-76.

Chorus, I. and J. Bartram (1999). Toxic Cyanobacteria in Water - A guide to their public health consequences, monitoring and management. Suffolk, St Edmundsbury Press.

Ciferri, O. (1983). "Spirulina, the edible microorganism." Microbiol Rev 47(4): 551-578.

Colovos, C. and T. O. Yeates (1993). "Verification of protein structures: patterns of nonbonded atomic interactions." Protein Sci 2(9): 1511-1519.

Crooks, G. E., G. Hon, J. M. Chandonia and S. E. Brenner (2004). "WebLogo: a sequence logo generator." Genome Res 14(6): 1188-1190.

Dangeard, P. (1940). "Sur une algue bleue alimentaire pour l'homme: Arthrospira platensis (Nordst.)." Gomont. Actes Soc. Linn. Boreaux Extr. Proces Verbaux 91: 39.

Davidson, A. L. and J. Chen (2004). "ATP-binding cassette transporters in bacteria." Annu Rev Biochem 73: 241-268.

Davidson, A. L., E. Dassa, C. Orelle and J. Chen (2008). "Structure, function, and evolution of bacterial ATP-binding cassette systems." Microbiol Mol Biol Rev 72(2): 317-364.

Desler, C., P. Suravajhala, M. Sanderhoff, M. Rasmussen and L. J. Rasmussen (2009). "In Silico screening for functional candidates amongst hypothetical proteins." BMC Bioinformatics 10: 289.

Devanathan, J., A. Selvam and N. Ramanathan (2016). "Optimization of biomass production of spirulina platensis in seawater medium." Life Science Archives 2(5): 708-716.

Dharmawardene, M. W., A. Haystead and W. D. Stewart (1973). "Glutamine synthetase of the nitrogen-fixing alga Anabaena cylindrica." Arch Mikrobiol 90(4): 281-295.

Dittmar, K. A., M. A. Sorensen, J. Elf, M. Ehrenberg and T. Pan (2005). "Selective charging of tRNA isoacceptors induced by amino-acid starvation." EMBO Rep 6(2): 151-157.

Doerks, T., V. Van Noort, P. Minguez and P. Bork (2012). "Annotation of the M. tuberculosis hypothetical orfeome: adding functional information to more than half of the uncharacterized proteins." PLoS One 7(4): e34302.

Doerks, T., C. Von Mering and P. Bork (2004). "Functional clues for hypothetical proteins based on genomic context analysis in prokaryotes." Nucleic Acids Res 32(21): 6321-6326.

Dong, H., L. Nilsson and C. G. Kurland (1996). "Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates." J Mol Biol 260(5): 649-663.

Dufresne, A., M. Salanoubat, F. Partensky, F. Artiguenave, I. M. Axmann, V. Barbe, S. Duprat, M. Y. Galperin, E. V. Koonin, F. Le Gall, K. S. Makarova, M. Ostrowski, S. Oztas, C. Robert, I. B. Rogozin, D. J. Scanlan, N. Tandeau de Marsac, J. Weissenbach, P. Wincker, Y. I. Wolf and W. R. Hess (2003). "Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome." Proc Natl Acad Sci U S A 100(17): 10020-10025.

Dvorak, P., A. Poulickova, P. Hasler, M. Belli, D. A. Casamatta and A. Papini (2015). "Species concepts and speciation factors in cyanobacteria, with connection to the problems of diversity and classification." Biodiversity and Conservation 24(4): 739-757.

Eisenberg, D., H. S. Gill, G. M. Pfluegl and S. H. Rotstein (2000). "Structure-function relationships of glutamine synthetases." Biochim Biophys Acta 1477(1-2): 122-145.

Eisenberg, D., R. Luthy and J. U. Bowie (1997). "VERIFY3D: assessment of protein models with three-dimensional profiles." Methods Enzymol 277: 396-404.

Elsaadi, O. and A. S. Cameron (1993). "Illness Associated with Blue-Green-Algae." Medical Journal of Australia 158(11): 792-793.

Fadi Aldehni, M., J. Sauer, C. Spielhaupter, R. Schmid and K. Forchhammer (2003). "Signal transduction protein P(II) is required for NtcA-regulated gene expression during nitrogen deprivation in the cyanobacterium Synechococcus elongatus strain PCC 7942." J Bacteriol 185(8): 2582-2591.

Falconer, I. R., A. M. Beresford and M. T. Runnegar (1983). "Evidence of liver damage by toxin from a bloom of the blue-green alga, Microcystis aeruginosa." Med J Aust 1(11): 511-514.

Farris, J. S. (1970). "Methods for Computing Wagner Trees." Systematic Biology 19(1): 83-92.

Fedyunin, I., L. Lehnhardt, N. Bohmer, P. Kaufmann, G. Zhang and Z. Ignatova (2012). "tRNA concentration fine tunes protein solubility." FEBS Letters 586(19): 3336-3340.

Felsenstein, J. (1985). "Confidence Limits on Phylogenies: An Approach Using the Bootstrap." Evolution 39(4): 783-791.

Ferjani, A., L. Mustardy, R. Sulpice, K. Marin, I. Suzuki, M. Hagemann and N. Murata (2003). "Glucosylglycerol, a compatible solute, sustains cell division under salt stress." Plant Physiol 131(4): 1628-1637.

Fiser, A. and A. Sali (2003). "Modeller: generation and refinement of homology-based protein structure models." Methods Enzymol 374: 461-491.

Fitch, W. M. (1971). "Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology." Systematic Biology 20(4): 406-416.

Fitch, W. M. and E. Margoliash (1967). "Construction of phylogenetic trees." Science 155(3760): 279-284.

Flores, E., J. E. Frias, L. M. Rubio and A. Herrero (2005). "Photosynthetic nitrate assimilation in cyanobacteria." Photosynthesis Research 83(2): 117-133.

Flores, E., J. E. Frías, L. M. Rubio and A. Herrero (2005). "Photosynthetic nitrate assimilation in cyanobacteria." Photosynthesis Research 83(2): 117-133.

Flores, E. and A. Herrero (2005). "Nitrogen assimilation and nitrogen control in cyanobacteria." Biochem Soc Trans 33(Pt 1): 164-167.

Forchhammer, K. (2004). "Global carbon/nitrogen control by PII signal transduction in cyanobacteria: from signals to targets." FEMS microbiology reviews 28(3): 319-333.

Forchhammer, K. and N. Tandeau de Marsac (1995). "Phosphorylation of the PII protein (glnB gene product) in the cyanobacterium Synechococcus sp. strain PCC 7942: analysis of in vitro kinase activity." J Bacteriol 177(20): 5812-5817.

Forde, B. G. (2000). "Nitrate transporters in plants: structure, function and regulation." Biochim Biophys Acta 1465(1-2): 219-235.

Forde, B. G. and P. J. Lea (2007). "Glutamate in plants: metabolism, regulation, and signalling." J Exp Bot 58(9): 2339-2358.

Frias, J. E., E. Flores and A. Herrero (1994). "Requirement of the regulatory protein NtcA for the expression of nitrogen assimilation and heterocyst development genes in the cyanobacterium Anabaena sp. PCC 7120." Mol Microbiol 14(4): 823-832.

Fujisawa, T., R. Narikawa, S. Okamoto, S. Ehira, H. Yoshimura, I. Suzuki, T. Masuda, M. Mochimaru, S. Takaichi, K. Awai, M. Sekine, H. Horikawa, I. Yashiro, S. Omata, H. Takarada, Y. Katano, H. Kosugi, S. Tanikawa, K. Ohmori, N. Sato, M. Ikeuchi, N. Fujita and M. Ohmori (2010). "Genomic structure of an economically important cyanobacterium, Arthrospira (Spirulina) platensis NIES-39." DNA Res 17(2): 85-103.

Gabbayazaria, R., M. Schonfeld, S. Telor, R. Messinger and E. Telor (1992). "Respiratory Activity in the Marine Cyanobacterium Spirulina-Subsalsa and Its Role in Salt Tolerance." Archives of Microbiology 157(2): 183-190.

Gadagkar, S. R., M. S. Rosenberg and S. Kumar (2005). "Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree." J Exp Zool B Mol Dev Evol 304(1): 64-74.

Gaget, V., M. Welker, R. Rippka and N. T. de Marsac (2015). "A polyphasic approach leading to the revision of the genus Planktothrix (Cyanobacteria) and its type species, P. agardhii, and proposal for integrating the emended valid botanical taxa, as well as three new species, Planktothrix paucivesiculata sp. nov.ICNP, and Planktothrix tepida sp. nov.ICNP, and Planktothrix serta sp. nov.ICNP, as genus and species names with nomenclatural standing under the ICNP." Syst Appl Microbiol 38(3): 141-158.

Galloway, J. N., F. J. Dentener, D. G. Capone, E. W. Boyer, R. W. Howarth, S. P. Seitzinger, G. P. Asner, C. C. Cleveland, P. A. Green, E. A. Holland, D. M. Karl, A. F. Michaels, J. H. Porter, A. R. Townsend and C. J. Vöosmarty (2004). "Nitrogen Cycles: Past, Present, and Future." Biogeochemistry 70(2): 153-226.

Galperin, M. Y. and E. V. Koonin (2004). "Conserved hypothetical' proteins: prioritization of targets for experimental study." Nucleic Acids Res 32(18): 5452-5463.

Galvan, A. and E. Fernandez (2001). "Eukaryotic nitrate and nitrite transporters." Cell Mol Life Sci 58(2): 225-233.

Garcia-Fernandez, J. M., N. T. de Marsac and J. Diez (2004). "Streamlined regulation and gene loss as adaptive mechanisms in Prochlorococcus for optimized nitrogen utilization in oligotrophic environments." Microbiol Mol Biol Rev 68(4): 630-638.

Gasteiger, E., A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel and A. Bairoch (2003). "ExPASy: The proteomics server for in-depth protein knowledge and analysis." Nucleic Acids Res 31(13): 3784-3788.

Geitler, L. (1925). Süsswasserflora Deutschlands, Österreichs und der Schweiz, . Jena, Fischer, G.

Gershwin, M. E. and A. Belay (2008). Spirulina in human nutrition and health. Boca Raton, CRC Press.

Gingold, H. and Y. Pilpel (2011). "Determinants of translation efficiency and accuracy." Mol Syst Biol 7: 481.

Ginsburg, A., J. Yeh, S. B. Hennig and M. D. Denton (1970). "Some effects of adenylylation on the biosynthetic properties of the glutamine synthetase from Escherichia coli." Biochemistry 9(3): 633-649.

Godia, F., J. Albiol, J. L. Montesinos, J. Perez, N. Creus, F. Cabello, X. Mengual, A. Montras and C. Lasseur (2002). "MELISSA: a loop of interconnected bioreactors to develop life support in space." J Biotechnol 99(3): 319-330.

Godoy Danesi, E. D., C. Oliveira Rangel-Yagui, S. Sato and J. C. Monteiro de Carvalho (2011). "Growth and content of spirulina platensis biomass chlorophyll cultivated at different values of light intensity and temperature using different nitrogen sources." Braz J Microbiol 42(1): 362-373.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin and S. G. Oliver (1996). "Life with 6000 genes." Science 274(5287): 546, 563-547.

Goffeau, A. and B. De Hertogh (2013). ABC Transporters. Encyclopedia of Biological Chemistry (Second Edition). W. J. Lennarz and M. D. Lane. Waltham, Academic Press. 1: 7-11.

Guerrero, M. G., J M Vega and M. Losada (1981). "The Assimilatory Nitrate-Reducing System and its Regulation." Annual Review of Plant Physiology 32(1): 169-204.

Guerrero, M. G., J. M. Vega and M. Losada (1981). "The Assimilatory Nitrate-Reducing System and its Regulation." Annual Review of Plant Physiology 32(1): 169-204.

Gupta, R. S. (2009). "Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades." Int J Syst Evol Microbiol 59(Pt 10): 2510-2526.

Habib, M. A. B., M. Parvin, T. C. Huntington and M. R. Hasan (2008). "A review on culture, production, and use of Spirulina as food for humans and feeds for domestic animals and fish."

Haft, D. H., B. J. Loftus, D. L. Richardson, F. Yang, J. A. Eisen, I. T. Paulsen and O. White (2001). "TIGRFAMs: a protein family resource for the functional identification of proteins." Nucleic Acids Research 29(1): 41-43.

Hamilton, T. L., D. A. Bryant and J. L. Macalady (2016). "The role of biology in planetary evolution: cyanobacterial primary production in low-oxygen Proterozoic oceans." Environ Microbiol 18(2): 325-340.

Heinrich, A., M. Maheswaran, U. Ruppert and K. Forchhammer (2004). "The Synechococcus elongatus P signal transduction protein controls arginine synthesis by complex formation with N-acetyl-L-glutamate kinase." Mol Microbiol 52(5): 1303-1314.

Hellriegel, H. and H. Wilfarth (1888). Untersuchungen über die Stickstoffnahrung der Gramineen und Leguminosen. Berlin, Buchdruckerei der "Post" Kayssler & Co.

Henderson, P. J. F. (1991). "Sugar transport proteins." Current Opinion in Structural Biology 1(4): 590-601.

Henson, B. J., S. M. Hesselbrock, L. E. Watson and S. R. Barnum (2004). "Molecular phylogeny of the heterocystous cyanobacteria (subsections IV and V) based on nifD." Int J Syst Evol Microbiol 54(Pt 2): 493-497.

Herrero, A., A. M. Muro-Pastor and E. Flores (2001). "Nitrogen control in cyanobacteria." J Bacteriol 183(2): 411-425.

Hille, R. (1996). "The Mononuclear Molybdenum Enzymes." Chem Rev 96(7): 2757-2816.

Hooft, R. W., G. Vriend, C. Sander and E. E. Abola (1996). "Errors in protein structures." Nature 381(6580): 272.

Horan, K., C. Jang, J. Bailey-Serres, R. Mittler, C. Shelton, J. F. Harper, J. K. Zhu, J. C. Cushman, M. Gollery and T. Girke (2008). "Annotating genes of known and unknown function by large-scale coexpression analysis." Plant Physiol 147(1): 41-57.

Humphrey, W., A. Dalke and K. Schulten (1996). "VMD: visual molecular dynamics." J Mol Graph 14(1): 33-38, 27-38.

Hunt, J. B., P. Z. Smyrniotis, A. Ginsburg and E. R. Stadtman (1975). "Metal ion requirement by glutamine synthetase of Escherichia coli in catalysis of gamma-glutamyl transfer." Arch Biochem Biophys 166(1): 102-124.

Ida, S. and B. Mikami (1983). "Purification and Characterization of Assimilatory Nitrate Reductase from the Cyanobacterium Plectonema boryanum." Plant and Cell Physiology 24(4): 649-658.

Irmler, A., S. Sanner, H. Dierks and K. Forchhammer (1997). "Dephosphorylation of the phosphoprotein P(II) in Synechococcus PCC 7942: identification of an ATP and 2-oxoglutarate-regulated phosphatase activity." Mol Microbiol 26(1): 81-90.

Ishii, S., S. Ikeda, K. Minamisawa and K. Senoo (2011). "Nitrogen cycling in rice paddy environments: past achievements and future challenges." Microbes Environ 26(4): 282-292.

Jackson, L. E., J. P. Schimel and M. K. Firestone (1989). "Short-Term Partitioning of Ammonium and Nitrate between Plants and Microbes in an Annual Grassland." Soil Biology & Biochemistry 21(3): 409-415.

Jha, P., A. Ali and N. Raghuram (2007). "Nitrate-Induction of Nitrate Reductase and its Inhibition by Nitrite and Ammonium Ions in Spirulina platensis." Physiol. Mol. Biol. Plants 13(2): 163-167.

Jha, P., A. Ali and N. Raghuram (2007). "Nitrate-Induction of Nitrate Reductase and its Inhibition by Nitrite and Ammonium Ions in Spirulina platensis." Physiology and Molecular Biology of Plants 13(2): 163-167.

Jiang, P. and A. J. Ninfa (1999). "Regulation of autophosphorylation of Escherichia coli nitrogen regulator II by the PII signal transduction protein." J Bacteriol 181(6): 1906-1911.

Jones, D. T., W. R. Taylor and J. M. Thornton (1992). "The rapid generation of mutation data matrices from protein sequences." Comput Appl Biosci 8(3): 275-282.

Jones, P. M. and A. M. George (2004). "The ABC transporter structure and mechanism: perspectives on recent research." Cell Mol Life Sci 61(6): 682-699.

Kameya, M., T. Ikeda, M. Nakamura, H. Arai, M. Ishii and Y. Igarashi (2007). "A novel ferredoxin-dependent glutamate synthase from the hydrogen-oxidizing chemoautotrophic bacterium Hydrogenobacter thermophilus TK-6." J Bacteriol 189(7): 2805-2812.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res 28(1): 27-30.

Kempf, B. and E. Bremer (1998). "Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments." Archives of Microbiology 170(5): 319-330.

Khademi, S., J. O'Connell, 3rd, J. Remis, Y. Robles-Colmenares, L. J. Miercke and R. M. Stroud (2004). "Mechanism of ammonia transport by Amt/MEP/Rh: structure of AmtB at 1.35 A." Science 305(5690): 1587-1594.

Khademi, S. and R. M. Stroud (2006). "The Amt/MEP/Rh family: structure of AmtB and the mechanism of ammonia gas conduction." Physiology (Bethesda) 21: 419-429.

Knaff, D. B. and M. Hirasawa (1991). "Ferredoxin-Dependent Chloroplast Enzymes." Biochimica Et Biophysica Acta 1056(2): 93-125.

Kobayashi, M., R. Rodriguez, C. Lara and T. Omata (1997). "Involvement of the C-terminal Domain of an ATP-binding Subunit in the Regulation of the ABC-type Nitrate/Nitrite Transporter of the Cyanobacterium Synechococcus sp. Strain PCC 7942." JOURNAL OF BIOLOGICAL CHEMISTRY 272(43): 27197-27201.

Kolb, A., S. Busby, H. Buc, S. Garges and S. Adhya (1993). "Transcriptional regulation by cAMP and its receptor protein." Annu Rev Biochem 62: 749-795.

Komarek, J., J. Kastovsky, J. Mares and J. R. Johansen (2014). "Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach." Preslia 86(4): 295-335.

Koropatkin, N. M., H. B. Pakrasi and T. J. Smith (2006). "Atomic structure of a nitrate-binding protein crucial for photosynthetic productivity." Proc Natl Acad Sci U S A 103(26): 9820-9825.

Krajewski, W. W., R. Collins, L. Holmberg-Schiavone, T. A. Jones, T. Karlberg and S. L. Mowbray (2008). "Crystal structures of mammalian glutamine synthetases illustrate substrate-induced conformational changes and provide opportunities for drug and herbicide design." J Mol Biol 375(1): 217-228.

Kudla, G., A. W. Murray, D. Tollervey and J. B. Plotkin (2009). "Coding-sequence determinants of gene expression in Escherichia coli." Science 324(5924): 255-258.

Kumada, Y., D. R. Benson, D. Hillemann, T. J. Hosted, D. A. Rochefort, C. J. Thompson, W. Wohlleben and Y. Tateno (1993). "Evolution of the glutamine synthetase gene, one of the oldest existing and functioning genes." Proc Natl Acad Sci U S A 90(7): 3009-3013.

Kumar, S., G. Stecher and K. Tamura (2016). "MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets." Mol Biol Evol 33(7): 1870-1874.

Kuypers, M. M. M., H. K. Marchant and B. Kartal (2018). "The microbial nitrogen-cycling network." Nat Rev Microbiol 16(5): 263-276.

Kyte, J. and R. F. Doolittle (1982). "A simple method for displaying the hydropathic character of a protein." J Mol Biol 157(1): 105-132.

Laalami, S. and H. Putzer (2011). "mRNA degradation and maturation in prokaryotes: the global players." Biomol Concepts 2(6): 491-506.

Latysheva, N., V. L. Junker, W. J. Palmer, G. A. Codd and D. Barker (2012). "The evolution of nitrogen fixation in cyanobacteria." Bioinformatics 28(5): 603-606.

Leonard, J. (1966). "The 1964–1965 Belgian Trans-Saharan expedition." Nature 209: 126–128.

Leonard, J. and P. Compare (1967). "Spirulina platensis (Gom) Geitl., algue bleue de grande valeur alimentaire par sa richeseen proteins." Bulletin Tu Jardin Botanique National Belgique 37: 3–23.

Letunic, I., R. R. Copley, S. Schmidt, F. D. Ciccarelli, T. Doerks, J. Schultz, C. P. Ponting and P. Bork (2004). "SMART 4.0: towards genomic data integration." Nucleic Acids Res 32(Database issue): D142-144.

Liaw, S. H. and D. Eisenberg (1994). "Structural model for the reaction mechanism of glutamine synthetase, based on five crystal structures of enzyme-substrate complexes." Biochemistry 33(3): 675-681.

Liaw, S. H., G. Jun and D. Eisenberg (1994). "Interactions of nucleotides with fully unadenylylated glutamine synthetase from Salmonella typhimurium." Biochemistry 33(37): 11184-11188.

Liaw, S. H., I. Kuo and D. Eisenberg (1995). "Discovery of the ammonium substrate site on glutamine synthetase, a third cation binding site." Protein Sci 4(11): 2358-2365.

Liaw, S. H., C. Pan and D. Eisenberg (1993). "Feedback inhibition of fully unadenylylated glutamine synthetase from Salmonella typhimurium by glycine, alanine, and serine." Proc Natl Acad Sci U S A 90(11): 4996-5000.

Lin, J. T. and V. Stewart (1998). "Nitrate assimilation by bacteria." Adv Microb Physiol 39: 1-30, 379.

Little, R., V. Colombo, A. Leech and R. Dixon (2002). "Direct interaction of the NifL regulatory protein with the GlnK signal transducer enables the Azotobacter vinelandii NifL-NifA regulatory system to respond to conditions replete for nitrogen." J Biol Chem 277(18): 15472-15481.

Little, R., F. Reyes-Ramirez, Y. Zhang, W. C. van Heeswijk and R. Dixon (2000). "Signal transduction to the Azotobacter vinelandii NIFL-NIFA regulatory system is influenced directly by interaction with 2-oxoglutarate and the PII regulatory protein." EMBO J 19(22): 6041-6050.

Lochab, S., P. A. Kumar and N. Raghuram (2014). "Molecular characterization of nitrate uptake and assimilatory pathway in Arthrospira platensis reveals nitrate induction and differential regulation." Arch Microbiol 196(6): 385-394.

Lochab, S., H. S. Oberoi, M. Gothwal, D. Abbey and N. Raghuram (2009). "Nitrate assimilatory enzymes of Spirulina (Arthospira) platensis are more thermotolerant than those of rice." Physiol Mol Biol Plants 15(3): 277-280.

Luo, F., Y. Yang, J. Zhong, H. Gao, L. Khan, D. K. Thompson and J. Zhou (2007). "Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory." BMC Bioinformatics 8: 299.

Luque, I., E. Flores and A. Herrero (1993). "Nitrite reductase gene from Synechococcus sp. PCC 7942: homology between cyanobacterial and higher-plant nitrite reductases." Plant Mol Biol 21(6): 1201-1205.

Luque, I., E. Flores and A. Herrero (1994). "Nitrate and Nitrite Transport in the Cyanobacterium Synechococcus Sp Pcc-7942 Are Mediated by the Same Permease." Biochimica Et Biophysica Acta-Bioenergetics 1184(2-3): 296-298.

Luque, I., M. F. Vazquez-Bermudez, J. Paz-Yepes, E. Flores and A. Herrero (2004). "In vivo activity of the nitrogen control transcription factor NtcA is subjected to metabolic regulation in Synechococcus sp. strain PCC 7942." FEMS Microbiology Letters 236(1): 47-52.

Mader, U., L. Zig, J. Kretschmer, G. Homuth and H. Putzer (2008). "mRNA processing by RNases J1 and J2 affects Bacillus subtilis gene expression on a global scale." Mol Microbiol 70(1): 183-196.

Madueño, F., M. A. Vega-Palas, E. Flores and A. Herrero (1988). "A cytoplasmic-membrane protein repressible by ammonium in Synechococcus R2: altered expression in nitrate-assimilation mutants." FEBS Letters 239(2): 289-291.

Maeda, S. I. and T. Omata (1997). "Substrate-binding lipoprotein of the cyanobacterium Synechococcus sp strain PCC 7942 involved in the transport of nitrate and nitrite." Journal of Biological Chemistry 272(5): 3036-3041.

Maheswaran, M., C. Urbanke and K. Forchhammer (2004). "Complex formation and catalytic activation by the PII signaling protein of N-acetyl-L-glutamate kinase from Synechococcus elongatus strain PCC 7942." J Biol Chem 279(53): 55202-55210.

Maia, L. B. and J. J. Moura (2015). "Nitrite reduction by molybdoenzymes: a new class of nitric oxide-forming nitrite reductases." J Biol Inorg Chem 20(2): 403-433.

Mann, N. H. and N. G. Carr (1992). Photosynthetic prokaryotes. New York; London, Springer US.

Manzano, C., P. Candau, C. Gomez-Moreno, A. M. Relimpio and M. Losada (1976). "Ferredoxin-dependent photosynthetic reduction of nitrate and nitrite by particles of Anacystis nidulans." Mol Cell Biochem 10(3): 161-169.

Marchler-Bauer, A. and S. H. Bryant (2004). "CD-Search: protein domain annotations on the fly." Nucleic Acids Res 32(Web Server issue): W327-331.

Marchler-Bauer, A., M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz, C. J. Lanczycki, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, Z. Wang, R. A. Yamashita, D. Zhang, C. Zheng and S. H. Bryant (2015). "CDD: NCBI's conserved domain database." Nucleic Acids Res 43(Database issue): D222-226.

Markou, G. (2015). "Fed-batch cultivation of Arthrospira and Chlorella in ammonia-rich wastewater: Optimization of nutrient removal and biomass production." Bioresour Technol 193: 35-41.

Markou, G., I. Chatzipavlidis and D. Georgakakis (2012). "Effects of phosphorus concentration and light intensity on the biomass composition of Arthrospira (Spirulina) platensis." World J Microbiol Biotechnol 28(8): 2661-2670.

Meeks, J. C., C. P. Wolk, J. Thomas, W. Lockau, P. W. Shaffer, S. M. Austin, W. S. Chien and A. Galonsky (1977). "The pathways of assimilation of 13NH4+ by the cyanobacterium, Anabaena cylindrica." J Biol Chem 252(21): 7894-7900.

Meinken, C., H. M. Blencke, H. Ludwig and J. Stulke (2003). "Expression of the glycolytic gapA operon in Bacillus subtilis: differential syntheses of proteins encoded by the operon." Microbiology 149(Pt 3): 751-761.

Merrick, M. J. and R. A. Edwards (1995). "Nitrogen control in bacteria." Microbiol Rev 59(4): 604-622.

Mikami, B. and S. Ida (1984). "Purification and properties of ferredoxin—nitrate reductase from the cyanobacterium Plectonema boryanum." Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology 791(3): 294-304.

Mohite, Y. S. and P. S. Wakte (2011). "Photosynthesis, growth and cell composition of Spirulina platensis (Arthrospira) under elevated atmospheric CO2 and nitrogen supplement." J. Algal Biomass Utln. 2(1): 77-94.

Najmudin, S., P. J. Gonzalez, J. Trincao, C. Coelho, A. Mukhopadhyay, N. M. Cerqueira, C. C. Romao, I. Moura, J. J. Moura, C. D. Brondino and M. J. Romao (2008). "Periplasmic nitrate reductase revisited: a sulfur atom completes the sixth coordination of the catalytic molybdenum." J Biol Inorg Chem 13(5): 737-753.

Nakano, S., M. Takahashi, A. Sakamoto, H. Morikawa and K. Katayanagi (2012). "Structure-function relationship of assimilatory nitrite reductases from the leaf and root of tobacco based on high-resolution structures." Protein Sci 21(3): 383-395.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol 48(3): 443-453.

Nelson, D. L. and M. M. Cox (2017). Lehninger principles of biochemistry. New York, W.H. Freeman and Company.

Oguchi, M., K. Otsubo, K. Nitta and S. Hatayama (1987). "Food production and gas exchange system using blue-green alga (Spirulina) for CELSS." Advances in space research : the official journal of the Committee on Space Research (COSPAR) 7(4): 7-10.

Ohashi, Y., W. Shi, N. Takatani, M. Aichi, S. Maeda, S. Watanabe, H. Yoshikawa and T. Omata (2011). "Regulation of nitrate assimilation in cyanobacteria." J Exp Bot 62(4): 1411-1424.

Omata, T. (1995). "Structure, function and regulation of the nitrate transport system of the cyanobacterium Synechococcus sp. PCC7942." Plant Cell Physiol 36(2): 207-213.

Omata, T., X. Andriesse and A. Hirano (1993). "Identification and characterization of a gene cluster involved in nitrate transport in the cyanobacterium Synechococcus sp. PCC7942." Mol Gen Genet 236(2-3): 193-202.

Omata, T., M. Ohmori, N. Arai and T. Ogawa (1989). "Genetically engineered mutant of the cyanobacterium Synechococcus PCC 7942 defective in nitrate transport." Proc Natl Acad Sci U S A 86(17): 6612-6616.

Padan, E., E. Bibi, M. Ito and T. A. Krulwich (2005). "Alkaline pH homeostasis in bacteria: new insights." Biochim Biophys Acta 1717(2): 67-88.

Padan, E., M. Venturi, Y. Gerchman and N. Dover (2001). "Na(+)/H(+) antiporters." Biochim Biophys Acta 1505(1): 144-157.

Padda, K. P., A. Puri and C. P. Chanway (2016). "Effect of GFP tagging of Paenibacillus polymyxa P2b-2R on its ability to promote growth of canola and tomato seedlings." Biology and Fertility of Soils 52(3): 377-387.

Pandey, J. P., N. Pathak and A. Tiwari (2010). "Standardization of pH and Light Intensity for the Biomass Production of Spirulina platensis." J. Algal Biomass Utln. 1(2): 93-102.

Pao, S. S., I. T. Paulsen and M. H. Saier, Jr. (1998). "Major facilitator superfamily." Microbiol Mol Biol Rev 62(1): 1-34.

Paz-Yepes, J., E. Flores and A. Herrero (2003). "Transcriptional effects of the signal transduction protein P(II) (glnB gene product) on NtcA-dependent genes in Synechococcus sp. PCC 7942." FEBS Letters 543(1-3): 42-46.

Peschek, G. A., C. Obinger, S. Fromwald and B. Bergman (1994). "Correlation between immuno-gold labels and activities of the cytochrome-c oxidase (aa3-type) in membranes of salt stressed cyanobactria." FEMS Microbiology Letters 124(3): 431-437.

Pesole, G., M. P. Bozzetti, C. Lanave, G. Preparata and C. Saccone (1991). "Glutamine synthetase gene evolution: a good molecular clock." Proc Natl Acad Sci U S A 88(2): 522-526.

Pisciotta, J. M., Y. Zou and I. V. Baskakov (2010). "Light-dependent electrogenic activity of cyanobacteria." PLoS One 5(5): e10821.

Plotkin, J. B. and G. Kudla (2011). "Synonymous but not the same: The causes and consequences of codon bias." Nat. Rev. Genet. 12(1): 32-42.

Ponce-Toledo, R. I., P. Deschamps, P. Lopez-Garcia, Y. Zivanovic, K. Benzerara and D. Moreira (2017). "An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids." Current Biology 27(3): 386-391.

Ponte-Sucre, A. (2009). ABC transporters in microorganisms : research, innovation and value as targets against drug resistance. Wymondham, Caister Academic Press.

Postgate, J. R. (1983). The fundamentals of nitrogen fixation. Cambridge, Cambridge University Press.

Puri, A., K. P. Padda and C. P. Chanway (2015). "Can a diazotrophic endophyte originally isolated from lodgepole pine colonize an agricultural crop (corn) and promote its growth?" Soil Biology & Biochemistry 89: 210-216.

Puri, A., K. P. Padda and C. P. Chanway (2016). "Evidence of nitrogen fixation and growth promotion in canola (Brassica napus L.) by an endophytic diazotroph Paenibacillus polymyxa P2b-2R." Biology and Fertility of Soils 52(1): 119-125.

Puri, A., K. P. Padda and C. P. Chanway (2016). "Seedling growth promotion and nitrogen fixation by a bacterial endophyte Paenibacillus polymyxa P2b-2R and its GFP derivative in corn in a long-term trial." Symbiosis 69(2): 123-129.

Quintana, N., F. Van der Kooy, M. D. Van de Rhee, G. P. Voshol and R. Verpoorte (2011). "Renewable energy from Cyanobacteria: energy production optimization by metabolic pathway engineering." Appl Microbiol Biotechnol 91(3): 471-490.

Rai, A. (2018). Handbook of Symbiotic Cyanobacteria. Boca Raton, CRC.

Ramachandran, G. N., C. Ramakrishnan and V. Sasisekharan (1963). "Stereochemistry of polypeptide chain configurations." J Mol Biol 7: 95-99.

Ramirez-Arcos, S., L. A. Fernandez-Herrero and J. Berenguer (1998). "A thermophilic nitrate reductase is responsible for the strain specific anaerobic growth of Thermus thermophilus HB8." Biochim Biophys Acta 1396(2): 215-227.

Raymond, J., O. Zhaxybayeva, J. P. Gogarten, S. Y. Gerdes and R. E. Blankenship (2002). "Whole-genome analysis of photosynthetic prokaryotes." Science 298(5598): 1616-1620.

Reed, R. H., L. J. Borowitzka, M. A. Mackay, J. A. Chudek, R. Foster, S. R. C. Warr, D. J. Moore and W. D. P. Stewart (1986). "Organic solute accumulation in osmotically stressed cyanobacteria." FEMS Microbiology Letters 39(1-2): 51-56.

Reitzer, L. (2003). "Nitrogen Assimilation and Global Regulation in Escherichia coli." Annual Review of Microbiology 57(1): 155-176.

Reitzer, L. (2003). "Nitrogen assimilation and global regulation in Escherichia coli." Annual Review of Microbiology 57: 155-176.

Rentsch, D., B. Hirner, E. Schmelzer and W. B. Frommer (1996). "Salt stress-induced proline transporters and salt stress-repressed broad specificity amino acid permeases identified by suppression of a yeast amino acid permease-targeting mutant." The Plant cell 8(8): 1437-1446.

Rice, P., I. Longden and A. Bleasby (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet 16(6): 276-277.

Richardson, D. J., B. C. Berks, D. A. Russell, S. Spiro and C. J. Taylor (2001). "Functional, biochemical and genetic diversity of prokaryotic nitrate reductases." Cell Mol Life Sci 58(2): 165-178.

Rippka, R. (1988). "Isolation and purification of cyanobacteria." Methods Enzymol 167: 3-27.

Ritchie, R. J. (1992). "Sodium-Transport and the Origin of the Membrane-Potential in the Cyanobacterium Synechococcus R-2 (Anacystis, Nidulans) Pcc-7942." Journal of Plant Physiology 139(3): 320-330.

Robertson, D. L. and A. Tartar (2006). "Evolution of glutamine synthetase in heterokonts: evidence for endosymbiotic gene transfer and the early evolution of photosynthesis." Mol Biol Evol 23(5): 1048-1055.

Rocap, G., F. W. Larimer, J. Lamerdin, S. Malfatti, P. Chain, N. A. Ahlgren, A. Arellano, M. Coleman, L. Hauser, W. R. Hess, Z. I. Johnson, M. Land, D. Lindell, A. F. Post, W. Regala, M. Shah, S. L. Shaw, C. Steglich, M. B. Sullivan, C. S. Ting, A. Tolonen, E. A. Webb, E. R. Zinser and S. W. Chisholm (2003). "Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation." Nature 424(6952): 1042-1047.

Rosano, G. L. and E. A. Ceccarelli (2009). "Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted Escherichia coli strain." Microb Cell Fact 8: 41.

Rowell, P., S. Enticott and W. D. P. Stewart (1977). "Glutamine synthetase and nitrogenase activity in the blue-green alga Anabaena cylindrica." New Phytologist 79(1): 41-54.

Rubio, L. M., E. Flores and A. Herrero (1998). "The narA locus of Synechococcus sp. strain PCC 7942 consists of a cluster of molybdopterin biosynthesis genes." J Bacteriol 180(5): 1200-1206.

Rubio, L. M., E. Flores and A. Herrero (1999). "Molybdopterin guanine dinucleotide cofactor in Synechococcus sp. nitrate reductase: identification of mobA and isolation of a putative moeB gene." FEBS Letters 462(3): 358-362.

Rubio, L. M., E. Flores and A. Herrero (2002). "Purification, cofactor analysis, and site-directed mutagenesis of Synechococcus ferredoxin-nitrate reductase." Photosynthesis Research 72(1): 13-26.

Rubio, L. M., A. Herrero and E. Flores (1996). "A cyanobacterial narB gene encodes a ferredoxin-dependent nitrate reductase." Plant Mol Biol 30(4): 845-850.

Ruppert, U., A. Irmler, N. Kloft and K. Forchhammer (2002). "The novel protein phosphatase PphA from Synechocystis PCC 6803 controls dephosphorylation of the signalling protein PII." Mol Microbiol 44(3): 855-864.

Saelices, L., D. Cascio, F. J. Florencio and M. I. Muro-Pastor "Crystal Structure of Glutamine Synthetase from Synechocystis sp. PCC 6803 ".

Saelices, L., R. Robles-Rengel, F. J. Florencio and M. I. Muro-Pastor (2015). "A core of three amino acids at the carboxyl-terminal region of glutamine synthetase defines its regulation in cyanobacteria." Molecular Microbiology 96(3): 483-496.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Mol Biol Evol 4(4): 406-425.

Sakamoto, T., K. Inoue-Sakamoto and D. A. Bryant (1999). "A novel nitrate/nitrite permease in the marine Cyanobacterium synechococcus sp. strain PCC 7002." J Bacteriol 181(23): 7363-7372.

Sazuka, T. (2003). "Proteomic analysis of the cyanobacterium Anabaena sp. strain PCC7120 with two-dimensional gel electrophoresis and amino-terminal sequencing." Photosynthesis Research 78(3): 279-291.

Schopf, J. W. (2014). "Geological evidence of oxygenic photosynthesis and the biotic response to the 2400-2200 ma "great oxidation event"." Biochemistry (Mosc) 79(3): 165-177.

Schopf, J. W. and B. M. Packer (1987). "Early Archean (3.3-billion to 3.5-billion-year-old) microfossils from Warrawoona Group, Australia." Science 237: 70-73.

Schuller, A., A. W. Slater, T. Norambuena, J. J. Cifuentes, L. I. Almonacid and F. Melo (2012). "Computer-based annotation of putative AraC/XylS-family transcription factors of known structure but unknown function." J Biomed Biotechnol 2012: 103132.

Seo, P. S. and A. Yokota (2003). "The phylogenetic relationships of cyanobacteria inferred from 16S rRNA, gyrB, rpoC1 and rpoD1 gene sequences." J Gen Appl Microbiol 49(3): 191-203.

Serrano, R. and A. Rodriguez-Navarro (2001). "Ion homeostasis during salt stress in plants." Curr Opin Cell Biol 13(4): 399-404.

Shahbabian, K., A. Jamalli, L. Zig and H. Putzer (2009). "RNase Y, a novel endoribonuclease, initiates riboswitch turnover in Bacillus subtilis." EMBO J 28(22): 3523-3533.

Shih, P. M., D. Wu, A. Latifi, S. D. Axen, D. P. Fewer, E. Talla, A. Calteau, F. Cai, N. Tandeau de Marsac, R. Rippka, M. Herdman, K. Sivonen, T. Coursin, T. Laurent, L. Goodwin, M. Nolan, K. W. Davenport, C. S. Han, E. M. Rubin, J. A. Eisen, T. Woyke, M. Gugger and C. A. Kerfeld (2013). "Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing." Proc Natl Acad Sci U S A 110(3): 1053-1058.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson and D. G. Higgins (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." Mol Syst Biol 7: 539.

Singh, P., S. S. Singh, M. Aboal and A. K. Mishra (2015). "Decoding cyanobacterial phylogeny and molecular evolution using an evonumeric approach." Protoplasma 252(2): 519-535.

Singh, P., S. S. Singh, J. Elster and A. K. Mishra (2013). "Molecular phylogeny, population genetics, and evolution of heterocystous cyanobacteria using nifH gene sequences." Protoplasma 250(3): 751-764.

Sivak, M. N., C. Lara, J. M. Romero, R. Rodriguez and M. G. Guerrero (1989). "Relationship between a 47-Kda Cytoplasmic Membrane Polypeptide and Nitrate Transport in Anacystis-Nidulans." Biochemical and Biophysical Research Communications 158(1): 257-262.

Smith, C. S., A. M. Weljie and G. B. Moorhead (2003). "Molecular properties of the putative nitrogen sensor PII from Arabidopsis thaliana." Plant J 33(2): 353-360.

Sokal, R. R. and C. D. Michener (1958). "A statistical method for evaluating systematic relationships." University of Kansas Science Bulletin 38(22): 1409–1438.

Soltis, P. S. and D. E. Soltis (2003). "Applying the bootstrap in phylogeny reconstruction." Statistical Science 18(2): 256-267.

Soong, F. S., E. Maynard, K. Kirke and C. Luke (1992). "Illness associated with blue-green algae." Med J Aust 156(1): 67.

Sorensen, M. A., J. Elf, E. Bouakaz, T. Tenson, S. Sanyal, G. R. Bjork and M. Ehrenberg (2005). "Over expression of a tRNA(Leu) isoacceptor changes charging pattern of leucine tRNAs and reveals new codon reading." J Mol Biol 354(1): 16-24.

Stewart, W. D. and P. Rowell (1975). "Effects of L-methionine-DL-sulphoximine on the assimilation of newly fixed NH3, acetylene reduction and heterocyst production in Anabaena cylindrica." Biochem Biophys Res Commun 65(3): 846-856.

Stolz, J. F. and P. Basu (2002). "Evolution of nitrate reductase: molecular and structural variations on a common function." Chembiochem 3(2-3): 198-206.

Suzuki, I., H. Kikuchi, S. Nakanishi, Y. Fujita, T. Sugiyama and T. Omata (1995). "A novel nitrite reductase gene from the cyanobacterium Plectonema boryanum." J Bacteriol 177(21): 6137-6143.

Szklarczyk, D., J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen and C. von Mering (2017). "The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible." Nucleic Acids Res 45(D1): D362-D368.

Tamura, K., M. Nei and S. Kumar (2004). "Prospects for inferring very large phylogenies by using the neighbor-joining method." Proceedings of the National Academy of Sciences of the United States of America 101(30): 11030-11035.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin and D. A. Natale (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics 4: 41.

Thajuddin, N. and G. Subramanian (2005). "Cyanobacterial biodiversity and potential applications in biotechnology." Current Science 89(1): 47-57.

Tiffany, L. H. (1968). Algae ; the grass of many waters. Spring field, Ill, Charles C. Thomas.

True, J. R. and S. B. Carroll (2002). "Gene co-option in physiological and morphological evolution." Annu Rev Cell Dev Biol 18: 53-80.

Turner, P. C., A. J. Gammie, K. Hollinrake and G. A. Codd (1990). "Pneumonia associated with contact with cyanobacteria." BMJ 300(6737): 1440-1441.

Van den Heuvel, R. H., D. Ferrari, R. T. Bossi, S. Ravasio, B. Curti, M. A. Vanoni, F. J. Florencio and A. Mattevi (2002). "Structural studies on the synchronization of catalytic centers in glutamate synthase." J Biol Chem 277(27): 24579-24583.

Van der Spoel, D., E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen (2005). "GROMACS: Fast, flexible, and free." Journal of Computational Chemistry 26(16): 1701-1718.

Van Eykelenburg, C. and A. Fuchs (1980). "Rapid reversible macromorphological changes in Spirulina platensis." Naturwissenschaften 67(4): 200-201.

Van Rooyen, J. M., V. R. Abratt, H. Belrhali and T. Sewell (2011). "Crystal structure of Type III glutamine synthetase: surprising reversal of the inter-ring interface." Structure 19(4): 471-483.

Vanoni, M. A. and B. Curti (1999). "Glutamate synthase: a complex iron-sulfur flavoprotein." Cell Mol Life Sci 55(4): 617-638.

Vega-Palas, M. A., E. Flores and A. Herrero (1992). "NtcA, a global nitrogen regulator from the cyanobacterium Synechococcus that belongs to the Crp family of bacterial regulators." Mol Microbiol 6(13): 1853-1859.

Vega-Palas, M. A., F. Madueno, A. Herrero and E. Flores (1990). "Identification and cloning of a regulatory gene for nitrogen assimilation in the cyanobacterium Synechococcus sp. strain PCC 7942." J Bacteriol 172(2): 643-647.

Volkl, P., R. Huber, E. Drobner, R. Rachel, S. Burggraf, A. Trincone and K. O. Stetter (1993). "Pyrobaculum aerophilum sp. nov., a novel nitrate-reducing hyperthermophilic archaeum." Appl Environ Microbiol 59(9): 2918-2926.

Vonshak, A. (1997). Spirulina platensis arthrospira: physiology, cell-biology and biotechnology. US and UK, Taylor & Francis.

Vonshak, A., R. Guy and M. Guy (1988). "The response of the filamentous cyanobacterium Spirulina platensis to salt stress." Arch. Microbiol. Archives of Microbiology 150(5): 417-420.

Walmsley, A. R., M. P. Barrett, F. Bringaud and G. W. Gould (1998). "Sugar transporters from bacteria, parasites and mammals: structure-activity relationships." Trends Biochem Sci 23(12): 476-481.

Wang, H., Y. Yang, W. Chen, L. Ding, P. Li, X. Zhao, X. Wang, A. Li and Q. Bao (2013). "Identification of differentially expressed proteins of Arthrospira (Spirulina) plantensis-YZ under salt-stress conditions by proteomics and qRT-PCR analysis." Proteome Sci 11(1): 6.

Wang, Q., H. Li and A. F. Post (2000). "Nitrate assimilation genes of the marine diazotrophic, filamentous cyanobacterium Trichodesmium sp. strain WH9601." J Bacteriol 182(6): 1764-1767.

Warr, S. R., R. H. Reed, J. A. Chudek, R. Foster and W. D. Stewart (1985). "Osmotic adjustment in Spirulina platensis." Planta 163(3): 424-429.

Warr, S. R. C., R. H. Reed and W. D. P. Stewart (1988). "The compatibility of osmotica in cyanobacteria." Plant Cell Environ Plant, Cell and Environment 11(2): 137-142.

Wiegand, C. and S. Pflugmacher (2005). "Ecotoxicological effects of selected cyanobacterial secondary metabolites: a short review." Toxicol Appl Pharmacol 203(3): 201-218.

Woese, C. R. (1987). "Bacterial evolution." Microbiol Rev 51(2): 221-271.

Wohlgemuth, S. E., T. E. Gorochowski and J. A. Roubos (2013). "Translational sensitivity of the Escherichia coli genome to fluctuating tRNA availability." Nucleic Acids Res 41(17): 8021-8033.

Wolk, C. P. (1973). "Physiology and cytological chemistry blue-green algae." Bacteriol Rev 37(1): 32-101.

Wolk, C. P., J. Thomas, P. W. Shaffer, S. M. Austin and A. Galonsky (1976). "Pathway of nitrogen metabolism after fixation of 13N-labeled nitrogen gas by the cyanobacterium, Anabaena cylindrica." J Biol Chem 251(16): 5027-5034.

Wu, Q. and V. Stewart (1998). "NasFED proteins mediate assimilatory nitrate and nitrite transport in Klebsiella oxytoca (pneumoniae) M5al." J Bacteriol 180(5): 1311-1322.

Wutipraditkul, N., R. Waditee, A. Incharoensakdi, T. Hibino, Y. Tanaka, T. Nakamura, M. Shikata, T. Takabe and T. Takabe (2005). "Halotolerant cyanobacterium Aphanothece halophytica contains NapA-type Na+/H+ antiporters with novel ion specificity that are involved in salt tolerance at alkaline pH." Appl Environ Microbiol 71(8): 4176-4184.

Xu, D. and Y. Zhang (2012). "Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field." Proteins 80(7): 1715-1735.

Yamashita, M. M., R. J. Almassy, C. A. Janson, D. Cascio and D. Eisenberg (1989). "Refined atomic model of glutamine synthetase at 3.5 A resolution." J Biol Chem 264(30): 17681-17690.

Zarembinski, T. I., L. W. Hung, H. J. Mueller-Dieckmann, K. K. Kim, H. Yokota, R. Kim and S. H. Kim (1998). "Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics." Proc Natl Acad Sci U S A 95(26): 15189-15193.

Zeng, M. T. and A. Vonshak (1998). "Adaptation of Spirulina platensis to salinity-stress." Comparative Biochemistry and Physiology a-Molecular & Integrative Physiology 120(1): 113-118.

Zhang, C. C., C. Z. Zhou, R. L. Burnap and L. Peng (2018). "Carbon/Nitrogen Metabolic Balance: Lessons from Cyanobacteria." Trends Plant Sci 23(12): 1116-1130.

Zhang, G., M. Hubalewska and Z. Ignatova (2009). "Transient ribosomal attenuation coordinates protein synthesis and co-translational folding." Nat Struct Mol Biol 16(3): 274-280.

Zhao, M. X., Y. L. Jiang, Y. X. He, Y. F. Chen, Y. B. Teng, Y. Chen, C. C. Zhang and C. Z. Zhou (2010). "Structural basis for the allosteric control of the global transcription factor NtcA by the nitrogen starvation signal 2-oxoglutarate." Proc Natl Acad Sci U S A 107(28): 12487-12492.

Zharkikh, A. and W. H. Li (1992). "Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock." Mol Biol Evol 9(6): 1119-1147.

Zhaxybayeva, O., J. P. Gogarten, R. L. Charlebois, W. F. Doolittle and R. T. Papke (2006). "Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events." Genome Res 16(9): 1099-1108.

Zheng, L., D. Kostrewa, S. Berneche, F. K. Winkler and X. D. Li (2004). "The mechanism of ammonia transport based on the crystal structure of AmtB of Escherichia coli." Proc Natl Acad Sci U S A 101(49): 17090-17095.

Zmasek, C. M. and S. R. Eddy (2001). "A simple algorithm to infer gene duplication and speciation events on a gene tree." Bioinformatics 17(9): 821-828.

Zouridis, H. and V. Hatzimanikatis (2008). "Effects of codon distributions and tRNA competition on protein translation." Biophys J 95(3): 1018-1033.

Zumft, W. G. (1997). "Cell biology and molecular basis of denitrification." Microbiol Mol Biol Rev 61(4): 533-616.

List of Publications

From Thesis:

Parva Sharma and Shibasish Chowdhury, "Evolutionary process of Glutamate Synthase Protein Family within the cyanobacteria: An *In-silico* Analysis" (Accepted for publication in IEEE Journal)

Parva Sharma and Shibasish Chowdhury, "*In-silico* analysis of nitrate assimilatory enzymes in cyanobacteria" (submitted) *Current Bioinformatics*

Parva Sharma and Shibasish Chowdhury, "Functional annotation of the un-annotated proteins of *Arthrospira platensis* NIES-39" (In Progress)

Papers presented in conferences:

Oral presentation at "International conference on Bioinformatics and Systems Biology", held at **IIIT, Allahabad** during 26-28 October, 2018

Poster presentation at "Annual Symposium of the Indian Biophysical Society", held at **IISER, Mohali** during 23-25 March, 2017

Poster presentation at "BITS Conference on Gene and Genome Regulation" held at **BITS-Pilani**, Pilani Campus during 18-20 Feb, 2016

Poster presentation at "International Conference on Proteomics from Discovery to Function", held at **IITB, Mumbai** during 7-9 Dec, 2014

Poster presentation at "International Conference on Biomolecular Forms and Functions", held at **IISC, Bangalore** during 8-11 Jan, 2013

From other projects:

Kanchan, S., **Sharma, P.** & Chowdhury, S., Evolution of endonuclease IV protein family: an *in silico* analysis, 3 Biotech (2019) 9: 168.

Biography of Prof. Shibasish Chowdhury

Prof. Shibasish Chowdhury obtained master's degree in physical chemistry from Calcutta University. Then, he shifted to biophysics and obtained PhD degree from Molecular Biophysics Unit (MBU) at Indian Institute of Science, Bangalore on "Computer modelling studies on G-rich unusual DNA structure". Subsequently, entered into protein folding field and worked as postdoctoral research fellow in the Department of Chemistry and Biochemistry, University of Delaware, USA for three years. Then, He joined department of Biological Science, BITS Pilani as lecturer in 2004, after that promoted to Assistant Professor (2006-2012) and then Associate Professor at the same department (2013-Till date). His broad research area is Protein folding, Modelling, Molecular evolution and Bioinformatics.

Biography of Parva Kumar Sharma

Mr. Parva Kumar Sharma has done his Bachelor of Science in Biotechnology from Kurukshetra University, Kurukshetra, Haryana and Master's Degree in Bioinformatics from CCS Haryana Agricultural University, Hisar, Haryana. He has also done Master of Engineering in Biotechnology from BITS Pilani, Goa Campus. Currently he is pursuing his doctoral thesis under the guidance of Prof. Shibasish Chowdhury, associate professor at Birla Institute of Technology and Science, Pilani, Pilani campus. He has also qualified CSIR-UGC National Eligibility Test (NET) for Lectureship in 'Life sciences' category in June 2012. During the period of Ph.D. research, he was awarded BITS Pilani Research Fellowship and Basic Science Research Fellowship from UGC, New Delhi. His research interest includes Bioinformatics data analysis, Molecular modeling, Structural bioinformatics, algorithm design and programming.