An *in silico* Study of Two Transcription Factors Controlling Diazotrophic Fates of the *Azolla* Major Cyanobiont *Trichormus azollae*

Bioinformatics and Biology Insights
Volume 14: 1–13

DOI: 10.1177/1177932220977490

Dilantha Gunawardana

Research Council, University of Sri Jayewardenepura, Nugegoda, Sri Lanka.

ABSTRACT: The cyanobiont *Trichormus azollae* lives symbiotically within fronds of the genus Azolla, and assimilates atmospheric nitrogen upon N-limitation, which earmarks this symbiosis to be a valuable biofertilizer in rice cultivation, among many other benefits that also include carbon sequestration. Therefore, studying the regulation of nitrogen fixation in *Trichormus azollae* is of great importance and benefit, especially the two topmost rungs of regulation, the NtcA and HetR transcription factors that are able to regulate the expression of myriads of downstream genes. Bioinformatics tools were used to zoom in on the NtcA and HetR transcription factors from *Trichormus azollae* to elaborate on what makes this particular cyanobiont different from other symbiotic as well as more distinct counterparts, in their commitment to nitrogen fixation. The utility of Azolla plants in tropical agriculture in particular merits the "top down N-regulation" by cyanobiont as a significant niche area of study, to make sense of superior N-fixing capabilities. The *Trichormus azollae* NtcA sequence was found as a phylogenetic outlier to horizontally infecting cyanobionts, which points to a distinct identity compared to symbiotic counterparts. There were borderline (60%-70%) levels of acceptable bootstrap support for the phylogenetic position of the Azolla cyanobiont's NtcA protein compared to other cyanobionts. Furthermore, the NtcA global nitrogen regulator in the Azolla cyanobiont has an extra cysteine at position 128, in addition to two other more conspicuous cysteines (positions, 157 and 164). A simulated homology model of the NtcA protein from *Trichormus azollae*, points to a single unique cysteine (Cysteine-128) as a key residue at the center of a lengthy C-helix, which forms a coiled-coil interface, through likely disulfide bond formation. Three cysteine (Cysteines: 128, 157, 164) architecture is exclusively found in *Trichormus azollae* and is absent in other cyanobacteria. A separate proline to alanine mutation in position 97—again exclusive to *Trichormus azollae*—appears to influence the flexibility of effector binding domain (EBD) to 2-oxoglutarate. The *Trichormus azollae* HetR sequence was found outside of horizontally-infecting cyanobiont sequences that formed a common clade, with the exception of the cyanobiont from the genus Cycas that formed one line of descent with the *Trichormus azollae* counterpart. Five (out of 6) serines predicted to be phosphorylated in the *Trichormus azollae* HetR sequence, are conserved in the *Nostoc punctiforme* counterpart, showcasing that phosphorylation is likley conserved in both vertically-transmitted and horizontally-acquired cyanobionts. A key Serine-127, within a conserved motif T**S**LT**S**, although conserved in heterocystous subsection IV and V cyanobacteria, are mutated in subsection III cyanobacteria that form trichomes but are unable to form heterocysts. I conclude that the NtcA protein from *Trichormus azollae* to be strategically divergent at specific amino acids that gives it an advantage in function as a 2-oxoglutarate-mediated transcription factor. The *Trichormus azollae* HetR transcription factor appears to possess parallel functionality to horizontally acquired counterparts. Especially Cysteine-128 in the NtcA transcription factor of the Azolla cyanobiont is an interesting proposition for future structure-function studies.

KEYWORDS: Nostoc, Trichormus, Anabaena, Sphaerospermopsis, NtcA transcription factor, HetR

RECEIVED: June 5, 2020. **ACCEPTED:** November 1, 2020.

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Dilantha Gunawardana, Research Council, University of Sri Jayewardenepura, Nugegoda, Sri Lanka. Email: dilantha@sci.sjp.ac.lk

Introduction

Azolla is a genus of water ferns, which are ubiquitous in parts of South and South East Asia due to the benefits it gifts paddy cultivation mainly through the fixation of atmospheric nitrogen, supplying ammonium ions to the host plant and surroundings. The nitrogen fixing microbe in Azolla, is a cyanobiont (a symbiotic cyanobacterium) designated *Trichormus azollae*, which is uncultivable *in vitro* and is inherited from the parent generation through vertical transmission, which dismisses the need for the reintroduction of the microbe from the environment.[1] It is inferred that the nitrogen fixation potential of the Azolla symbiotic system, varies from 30 to 60 kg/hectare of nitrogen per year, which makes it a powerhouse in nitrogen fixation.[2] It is also suggested that Azolla can produce the equivalent of 1800 kg of urea per hectare per year.[3]

Trichormus azollae, the cyanobiont, is controversial in its taxonomy.[4] The cyanobiont has been called by three generic names this far; *Anabaena*, the earliest, *Nostoc*, the most common, and *Trichormus*, the most recent and the least common.[4] Although a single genus name has not be consolidated, it is fair to say that there are certain features of the cyanobiont that are of interest, as a N-fixing microorganism. There is erosion of the cyanobiont's genome, which outside of nitrogen fixation and a few other key functions, relegates the cyanobiont to a committed and obligate relationship with the symbiotic host.[5] Furthermore, the photosynthesis potential of the cyanobiont too has been reduced due to its dependence on the host Azolla plant for photosynthate.[6] In addition, the nitrogen fixation potential is crucial for the rapidly dividing fern and the surrounding ecosystem, which makes this symbiosis one of strong utility in sustainable agriculture.

A newer paradigm is presented by Azolla due to the plant's potential as a voracious carbon sink—the Rubisco Carboxylase is the most common enzyme in plants -, which can be utilized

in abundance due to the strong nitrogen fixation performed by the cyanobiont.[7] I showed in an earlier publication that *Trichormus-Azolla* symbiosis is a sound future bio-technological ally in a prospective "field-based" setting, due to the absence of nitrous oxide and methane emitting enzymes in the cyanobiont's proteome.[7] A third benefit that can be reaped from Azolla, is its potent ability to quench harmful heavy metals such as lead, chromium and to a lesser extent cadmium.[8] Again, phytoremediation is dependent on proteins such as "sequestering" metallothioneins, and therefore, nitrogen fixation by the cyanobiont, once again, is of strong importance.[8]

In large, the many benefits to be reaped from the *Azolla-Trichormus azollae* system, is protein-dependent which makes the nitrogen fixation engine of the cyanobiont, a valuable cog of contemporary research. The key protein in nitrogen management in *Trichormus azollae*—and in many other cyanobacteria—is the NtcA transcription factor.[9,10] The NtcA transcription factor belongs to the cAMP (CRP) transcription factor family of proteins and is crucial for the downstream activity of a vast nitrogen metabolism network, comprising of hundreds of genes.[11] In fact, 11 of the 13 residues that forms contacts with DNA in CRP transcription factors are conserved or replaced with a conservative residue in the NtcA proteins. The compound 2-oxoglutarate is thought to activate the functional confirmation of the NtcA protein and in its absence, it is assumed the transcription factor to be in an inactive state. However, even in the inactive state, there is some form of DNA binding, while activation by 2-oxoglutarate enhances the binding affinities.[12] Furthermore, the NtcA protein in cyanobacteria is thought to aid in RNA polymerase recruitment and not merely in DNA binding-based regulation of gene expression.[11]

HetR, the next in line cog to the NtcA transcription factor, has been shown to be responsive to the N-state of the cell, downstream of 2-oxoglutarate, which is sensed allosterically by the NtcA sensor protein. Some empirical work has been performed on HetR transcription factors which are key for the formation of heterocysts. In particular, serines have been shown to be key for both signal transduction by phosphorylation, as well as for nucleophile function to enzymatically effectuate catalysis.[13] In heterocystous cyanobacteria, a key motif that reads T**S**LT**S** is highly conserved. A recent study demonstrated an upstream kinase (Pnk22) that is capable of phosphorylation of HetR proteins.[13]

In symbiotic systems, there can be regulation of expression of selective genes in a symbiont, as compared to the stage of development of the host. In the symbiotic strain *Nostoc punctiforme*, the expression of the *ntcA* gene after a single infection of the host *Gunnera* plants, has been found to be minor in early developmental stages of the host, increasing during the middle stages of development of host and finally receding in the older *Gunnera* plants (Wang, Ekman et al., 2004). So did the HetR protein, which too produced the most mRNA during the middle stages (stage 3) of development of the host plant (Wang, Ekman et al., 2004).[14] It is now known that the transcription profiles of middle stages of development of an organism are predominantly due to ancient gene families (de Mendoza, Sebe-Pedros et al., 2013).[15] The NtcA transcription family is an ancient and conspicuous protein encoded by N-fixing cyanobacterial genomes and can be compared due to a longer window of gradual evolution and due to the richness of species that are covered by the protein. NtcA transcription factor has been shown to bind to 2424 DNA elements, of which 2153 are genes (Picossi, Flores et al., 2014).[16]

This *in silico* study explores the phylogeny of the Azolla cyanobiont, *Trichormus azollae*, using NtcA proteins due to its essential "top of the pyramid" location and the functional significance of key residues that are distinct for this protein. I too explore the HetR transcription factors in terms of their phylogeny, the putative serine residues that are likely to be phosphorylated to relay downstream the signal for cell differentiation in heterocyst formation and subsequent nitrogen fixation. The landscape of nitrogen-centered functions is discussed using the structures of NtcA and HetR proteins as foci. This study advances the field of knowledge on this symbiont residing inside a water fern, which is distinct from all other plants, in harboring a vertically-perpetuating cyanobiont.

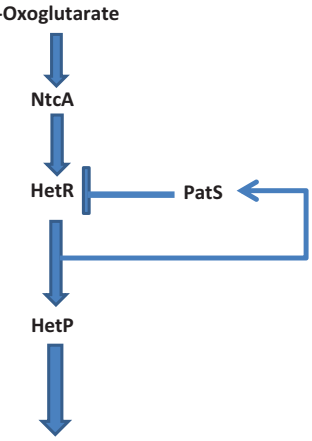
Results and Discussion

Phylogeny of NtcA transcription factors in cyanobacteria and among cyanobionts

The hierarchy of the mechanism of heterocyst differentiation begins with the sensing of nitrogen depletion/starvation and the commencement of the downward cascade commencing with the global nitrogen regulator, the NtcA transcription factor (Figure 1). It has been demonstrated in the strain *Anabaena* sp. PCC 7120, that heterocyst differentiation is clocked as below: induction of differentiation (0-2 hours). Formation of pattern along the trichome (2-9 hours), commitment to a differentiated state (9-13 hours), morphological changes (13-24 hours).[17] In 48 hours, there is a full fledged heterocyst along a trichome previously composed of a string of vegetative cells[18] to commence nitrogen fixation.

The phylogenetic reconstruction was performed using three other genera that all showed strong sequence identity at 100% coverage with the *Trichormus azollae* NtcA protein. The phylogenetic trees (Figure 2A) based on four proximal genera, were inconclusive for phylogenetic inferences due to low bootstrap support. The difficulty in arriving at conclusive phylogeny compared to neighbors, appears to be curtailed by the lack of sufficient divergence of sequences and insufficient sites (of mutations) for study.

Anabaena cylindrica, forms a monophyletic cluster with *Trichormus azollae*, when the homocitrate synthase protein sequences were assessed for their phylogenetic relationships.[19] Here in this study, *Anabaena cylindrica*, forms a neighboring clade to a unitary *Trichormus azollae*, in relation to NtcA protein phylogeny, and shares the specific clade with a member of the genus *Sphaerospermopsis*, although not backed by



Host (Genus)	Cyanobiont	Age of Symbiosis
Azolla	*Trichormus azollae*	~90 Mya
Gunnera	*Nostoc punctiforme*	~115 Mya
Cycas	*Nostoc cycadae*	~265-290 Mya
Anthoceros	*Nostoc punctiforme*	~470 Mya

Figure 1. (A): The top-down hierarchy of NtcA dependent differentiation of heterocysts. (B) Cyanobionts and their putative (maxiumum) symbiotic age (Warshan et al., 2018).[20]

strong bootstrap support. This unsubstantiated finding (due to inconclusive sampling support) is supplementary and supportive of recent single-gene and whole-genome phylogeny studies. Interestingly, the genus *Sphaerospermopsis* (specifically *Sphaerospermopsis aphanizomenoides* BCCUSP55) has been shown to form a single two-member monophyletic clade with *Trichormus azollae*, with 64% bootstrap support between nodal tips, when the rbcL-rbcX molecular marker was employed.[21] Still, when whole genome phylogeny was inferred for cyanobacteria, the *Trichormus azollae* genome formed a monophyletic clade only with the genome of *Sphaerospermopsis aphanizomenoides* BCCUSP55 (Warshan et al., 2018). The genus *Sphaerospermopsis* forms coiled and straight filaments and, in some cases, has been originally thought as *Anabaena*, exclusively from morphology, and later changed in nomenclature based on molecular data gathering exercises.

Next, I searched individually for NtcA proteins in the NCBI protein database from known cyanobionts and constructed a Maximum Likelihood phylogenetic tree from the downloaded sequences (10 in number) with bootstrap support from 500 pseudo-replications. The phylogenetic tree of the cyanobiont NtcA proteins demonstrated that the *Trichormus azollae* NtcA protein formed a lone member outside of all other cyanobionts that clustered together, further dividing into smaller daughter clades. The bootstrap support for the position of the *Trichormus azollae* NtcA protein, is $>60\%$ but $<70\%$, which tells us that they are likely to be accurate than inaccurate. Why the NtcA protein of *Trichormus azollae* is closer to the free-living species (Figure 2A) and falls outside of the core "cyanobionts" clade (Figure 2B), does present an interesting biological conundrum. In this, the structure-function relationship of the NtcA protein of *Nostoc azollae* becomes crucial to attest to the changes of sequence that can contribute to its role in N-regulation.

In relation to cyanobionts, the *Azolla-Trichormus* relationship is the youngest (~90 MYA) and except for the *Gunnera-Nostoc* symbiosis dating to ~115 MYA (Warshan et al., 2018) are ancient symbioses that have coevolved during the establishment of plant symbioses, namely those of gymnosperms and bryophytes (Figure 1B) (It should be noted that due to the proximity of 90 MYA and 115 MYA, they could very well be contemporary events in age). Perhaps, it is the shorter evolutionary history of the NtcA transcription factor of *Trichormus azollae* that clusters it closer to the free-living species of cyanobacteria and distant from other cyanobionts (Figure 1B). *Trichormus azollae* has been known to produce more frequently spaced heterocysts in N-starved conditions and is a powerhouse in relation to nitrogen fixation.

Functional insights to the NtcA transcription factor in Trichormus azollae

A few non-conserved positions in the *Trichormus azollae* NtcA sequence were observed when aligned using ClustalW. One such unique mutation, I infer to be of strong functional significance (Figure 3). The NtcA transcription factor from the *Trichormus azollae* cyanobiont possesses three cysteines, which is one more in number than all auxiliary sequences from cyanobacteria. While the cysteine in position 157 is 100% conserved, a majority of sequences used for the sequence alignment possesses the cysteine in position 164. However, the cysteine at position 128, is exclusively found in *Trichormus azollae* (Figures 3 and 4). It has been suggested that the presence of two cysteines (Position 157 and Position 164) in cyanobacteria such as *Anabaena* sp. PCC 7120 (which too are present in *Trichormus azollae* in the same positions) is an indicator of intra-molecular disulfide bond formation, although this hypothesis has been proven to be inaccurate.[22]

Other biological pathways outside of nitrogen metabolism, such as the light-dependent keto caretenoid pathway, are dependent on the NtcA transcription factor.[23] Therefore, the NtcA transcription factor acts as a universal "manager," for the regulation of many pathways, especially those involving the element nitrogen. The DNA binding helix-turn-helix motif is found conserved between residues 174 to 195 in the NtcA global nitrogen regulator protein family[22] and is known to allow optimal spacing for DNA binding by shifting the two helices apart, upon activation by 2-oxoglutarate[12] and this relay of function from the sensory N-terminus to the effector C-terminus is thought to be performed by the central C-helix.

Cysteines have features that are crucial from a functional perspective. Cysteines possess thiol/sulfhydryl groups (the only amino acid to utilize such a group), are capable of forming disulfide bonds using two cysteine residues, are found as a highly conserved residue in protein sequences, forms clusters in close proximity, have high metal binding affinities, while

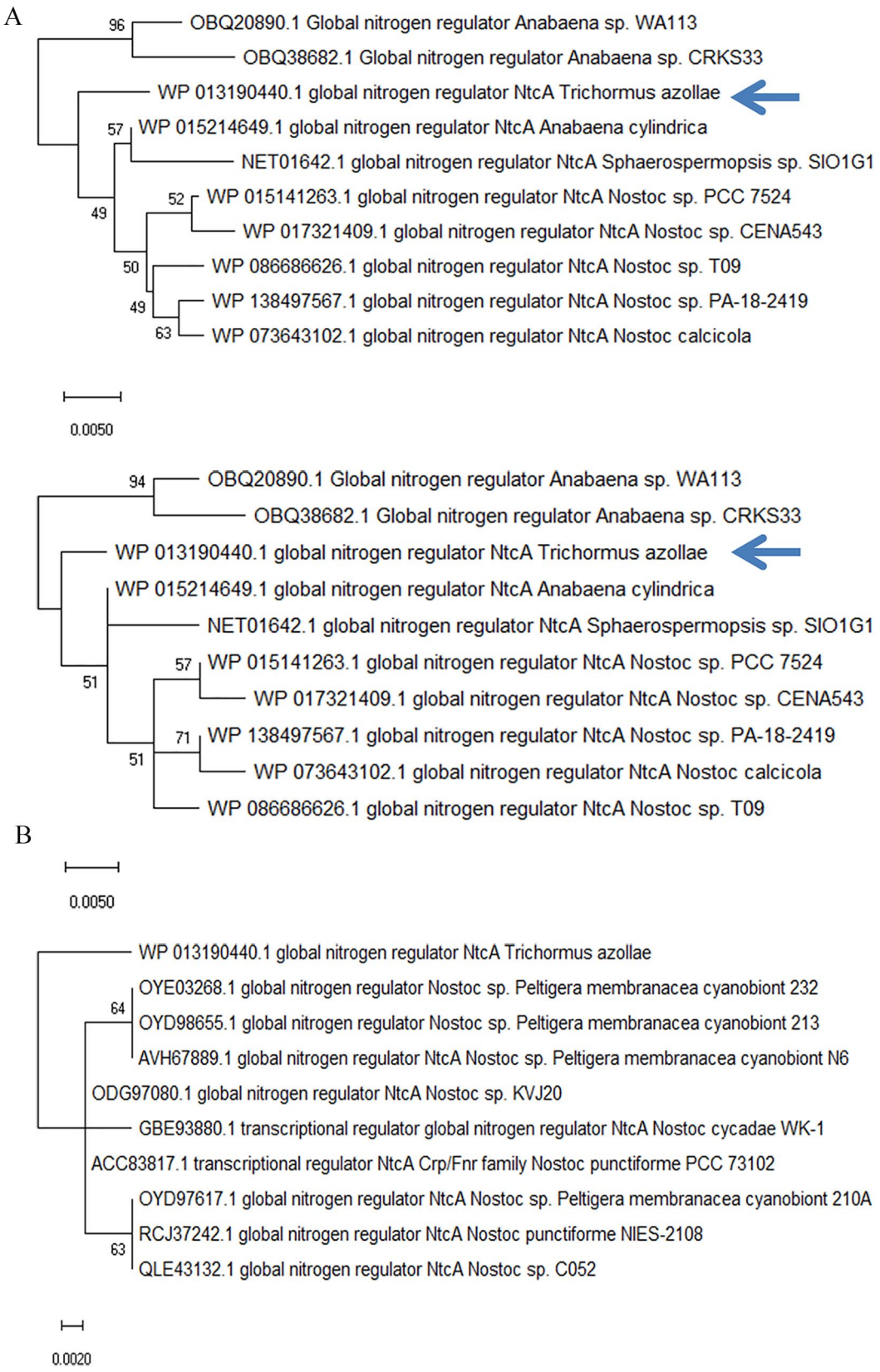


Figure 2. (A): Phylogeny of NtcA transcription factors in selected cyanobacteria: The amino acid sequences of NtcA proteins from four genera that all showed high sequence homology at 100% sequence coverage to the *Trichormus azollae* NtcA sequence (WP_013190440), were first aligned using the ClustalW algorithm using MEGA version X and the phylogenetic reconstruction performed using both the Neighbor Joining method (top) and Maximum Likelihood (bottom) method with support from 1000 bootstrap replications. (B): Phylogeny of NtcA transcription factors from cyanobionts: The amino acid sequences of 10 cyanobionts were first aligned and the phylogenetic tree constructed using the Maximum Likelihood method, with bootstrap support from 500 replications.

Figure 3. Multiple sequence alignment of NtcA transcription factors from cyanobacteria. WP_013190440 is the *Trichormus azollae* NtcA sequence, which is the only protein sequence to have a third cysteine at position 128. Other cyanobacterial NtcA proteins have only two cysteines in primary sequence.

attracting controversy on its hydrophobic/non-hydrophobic nature.[24] Such properties in cysteines confine them to be largely irreplaceable. The conversion of a serine to a cysteine (Figure 3), is no wobble/third position change, and requires a two nucleotide conversion (Figure 4E) which demonstrates that the conversion of a serine to a cysteine is most likely a functional adaptation. According to the Grantham's distance based on composition, polarity and molecular volume of an amino acid and Miyata's distance, which is based on volume and polarity (Table 1), serine to cysteine transformations (through non-synonymous substitution of codons) can be termed significant due to their physiochemical distance between cysteine and serine, which suggests that the reason for the change is functional or operational. Cysteines too are rare in helices and are predominantly found as part of beta sheets,[25] which too points to the cysteine here as performing a key/discrete function.

Cys-128 is found in the lengthy central C-helix that forms the tight coiled-coil interface between the reacting monomers. A comparison of the homology model of the *Trichormus azollae* NtcA protein against that of *Anabaena cylindrica* homology model, showcases that the central interface is narrower in the former, while in the latter, a broader gap is found between the dimer partners (Figure 4B and C). Upon binding of 2-oxoglutarate, a crucial twisting of the central C-helix, from a 17 degree angle to a 23 degree angle, helps form a coiled coil structure (Figure 4D), a stable dimer with a significantly larger binding interface (approximately 2000 Å^2)[12] The surrounding region corresponding to the Cys-128, is termed the helical bridge or the C-helix, which is able to upon N-terminal 2-oxoglutarate binding, change its conformation and bind tighter than in the inactive-conformation.[12] Reformation of hydrogen bonds of key residues in the paired helices of the dimer (with 2-oxoglutarate), in particular hydrogen bonds donated by arginines and glutamates, are

Figure 4. *(Continued)*

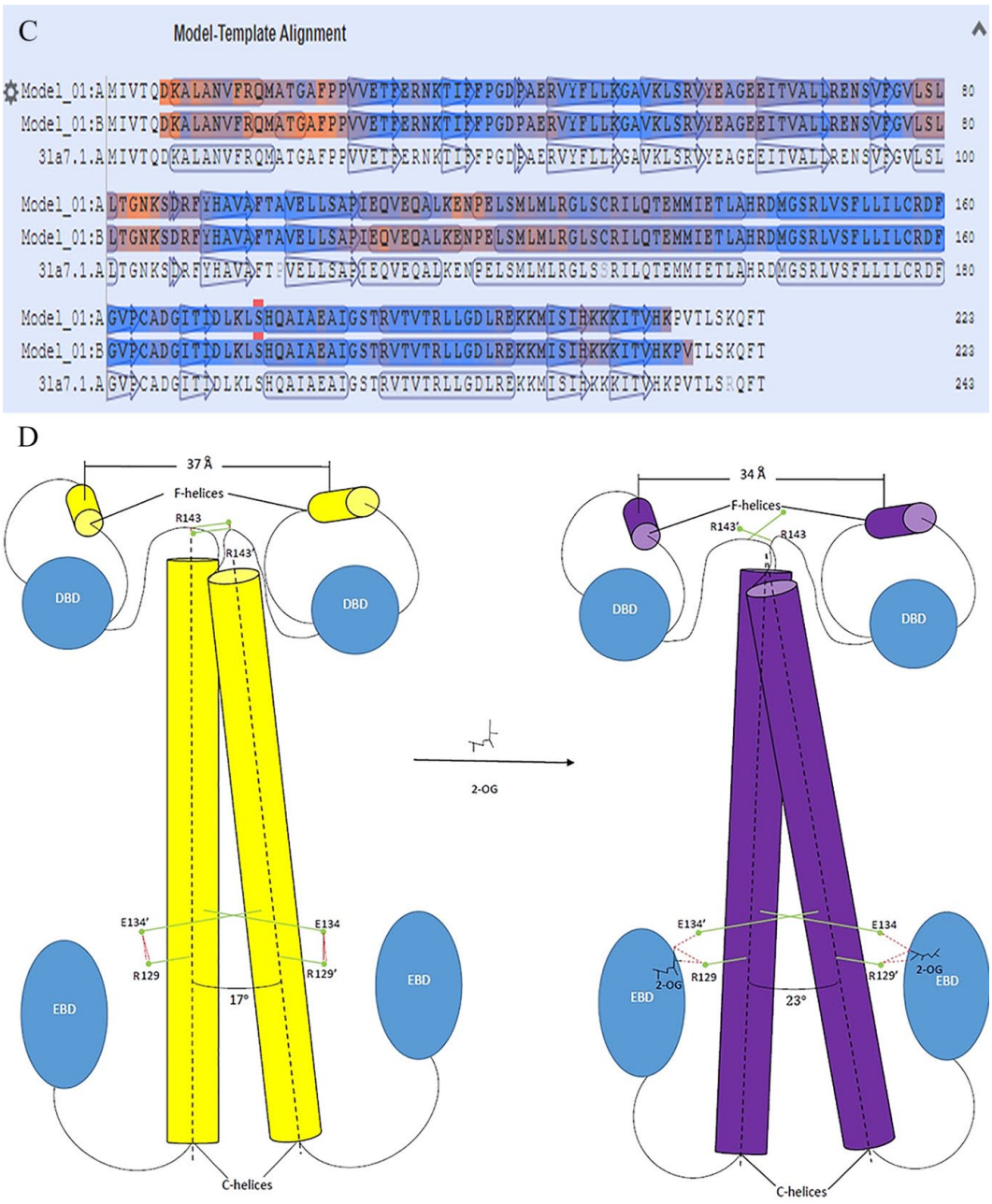


Figure 4. *(Continued)*

Figure 4. (A): Output from the secondary structure prediction tool PSIPRED 4.0 for the NtcA protein from *Trichormus azollae* showing helices, beta strands and loops/coils. The color-coded specifics are shown below the panel. (B): The tertiary structure (homology model) of the NtcA protein of the Azolla cyanobiont (*Trichormus azollae*) modeled using the SWISS-MODEL web server (see methods section), and compared to the homology model of the *Anabaena cylindrica* structural homolog. The dimer interfaces are shown by a bracket in red. The template ID is 31a7.1.A. (C): The alignment of the model (chains A and B of the homodimer) with the reference structure (3la7.1.A) as shown in SWISS-MODEL. The secondary structural elements are shown below the sequences. (D): Illustration of the mechanism of formation of a coiled coil structure by the NtcA protein dimer, using the central C-helix as the coil. (EBD—Effector Binding Domain; DBD—DNA Binding Domain). The angle of one monomer against the other, expands from a 17 degree angle to a 23 degree angle upon 2-oxoglutarate binding. The hydrogen bonds and the effector binding too are shown in the model. This was illustrated as demonstrated in Zhao et al.[12] (E): Stretch of sequence of NtcA protein of *Trichormus azollae* aligned against those of *Anabaena variabilis* (U89516.1) belonging to Subsection IV, *Cyanothece* ATCC51142 (U80855.1) in subsection I and *Microcystis aeruginosa* PCC 7806 (EU402445.1) which too is grouped in subsection I. Two changes in amino acids composition only found in *Trichormus azollae* sequence is shown in boxes, against those of cyanobacterial counterparts. The asterisks below indicate the number of nucleotide changes/level of preservation. Subsection I members are strongly divergent from subsection IV.

Table 1. Empirical amino acid substitution pattern from DAMBE,[26] based on sequence pairs comprising 10 cyanobiont sequences in total. Amino acid dissimilarity indices are Grantham's distance,[27] Miyata's distance[28] and neighbor-based distance.[25] Ones in bold are from *Trichormus azollae* NtcA protein sequence, against other symbiotic counterparts.

AA1—AA2	NUMBER	G	M	N
Leu—Gln	21	112	2.696	114.9
Lys—Arg	9	26	0.397	70.5
Pro—Ala	**9**	**27**	**0.064**	**56.6**
Ser—Ala	21	99	0.509	41.0
Ser—Cys	**9**	**112**	**1.836**	**47.1**
Thr—Ser	9	58	0.885	23.6

Abbreviations: G, Grantham's distance; M, Miyata's distance; N, neighbor-based distance.

Table 2. Predicted disulfide bonds and their probabilities based on scores, when the *Trichormus azollae* NtcA protein sequence was checked using the DiANNA server (http://clavius.bc.edu/~clotelab/DiANNA/).[29] No putative disulfide bond formations were predicted.

DISULFIDE BOND SCORES			
CYSTEINE SEQUENCE POSITION	DISTANCE	BOND	SCORE
128-157	29	LRGLSCRILQT-FLLILCRDFGV	0.01064
128-164	36	LRGLSCRILQT-DFGVPCADGIT	0.01062
157-164	7	FLLILCRDFGV-DFGVPCADGIT	0.01073

termed significant in fact, Arginine-129 and Glutamate-134 bind 2-oxoglutarate directly, immediately sensing the binding of the ligand. In particular, Arginines 129 (in the C-helix) and at position 143 (at the hinge region), and Glutamates 134 and 135 from the C-helix, are key resiudes for conformational adaptation and relaying the conformational changes to the DNA-binding domain upon 2-oxoglutarate binding.[12]

The central location of Cysteine-128 in the C-helix and the inter-molecular propensity of cysteines to form disulfide bonds/bridges, suggest that the two Cysteines at position 128, are involved in the formation of an inter-molecular disulfide-bond. The dimer would be tighter in binding compared to other NtcA proteins that are absent of a third cysteine in the C-helix, and thereby are reliant solely on inter-molecular hydrogen bonds (Figure 4B). In fact, Cysteine-128 is strategically placed at the center of the C-helix, 13 residues immediately adjacent to the right and 12 residues found on the left (Figure 4A and C). When the putative intra-molecular disulfide bonds were predicted for the *Trichormus azollae* NtcA protein, not a single disulfide bond was predicted by the prediction service (Table 2), which too supports the theory that the inter-molecular disulfide bond formation may likely assist in the dimerization of the NtcA monomers in the absence of intra-molecular disulfide bonds.

Coiled coil domains are widespread as dimerization interfaces performing key regulatory functions. Examples here are transcription factors such as C-fos and C-jun proteins.[30,31] Intermolecular dimerization in the presence of a cysteine disulfide bridge is also thought to enhance the thermal stability of the protein. Structural biology in tandem with gel mobility shift assays and site-directed mutagenesis studies are required to demonstrate such hypotheses.

Another mutation (position 97) of NtcA sequence of *Trichormus azollae*, encompasses a proline to alanine transformation (Figure 4E) which is not found in counterparts from other cyanobacterial divisions, namely *Anabaena variabilis* (U89516.1) belonging to subsection IV, unicellular *Cyanothece* ATCC51142 (U80855.1) in subsection I that has 34 *nif* genes (the most in cyanobacteria), *Microcystis aeruginosa* PCC 7806 (EU402445.1) in subsection I that is capable of cyanotoxin production. Prolines are residues that due to a dearth in hydrogen bond donation capacities are found mostly in loops/turns and not in helices or beta sheets. The proline-97 forces the FTA tripeptide sequence between two beta strands (Figure 4C) to be rigid and the replacement with an alanine in the *Trichormus azollae* sequence, changes the rigidity to a more flexible structure. I infer from the tertiary structure that the effector binding domain of the NtcA protein (Figure 4D) where this proline to alanine transformation is found and its close proximity to the 2-oxoglutarate binding pocket (Figure 4B), makes this mutation one that influences structural flexibility accompanying effector binding and subsequent dimerization.

Phylogeny of HetR transcription factors in cyanobionts

When 10 cyanobiont HetR sequences were used to construct a Maximum Likelihood phylogenetic tree, I found that the *Trichormus azollae* and *Nostoc cycadaea* wk-1 HetR sequences formed on a distinct lineage distant from other cyanobionts that formed a collective cluster that further trifurcated into three daughter clades. Bootstrap support for the position of the HetR sequence *of Trichormus azollae* is strong, with 62%-100% bootstrap support (Figure 5). Again, I am not able to distinguish the cyanobionts from Azolla fronds, and those from Cycad coralloid root nodules, in relation to their functional significance, although it is known that the Cycas counterpart needs infection of the root system while in *Trichormus azollae*, it is a case of vertical transmission, which needs no infection from the surrounding environment. Furthermore, the genus Cycas symbionts are evolutionary older (~260-290 MYA) compared to the Azolla cyanobiont which does pose key questions, at their relative mutation rates and evolutionary pathways. Still *Trichormus azollae* is an obligate symbiont and has been only subjected to symbiotic pressures for ~90 MYA, while the Cycas counterpart is facultative, which suggests that the evolutionary pressures to be different between the two cyanobionts and consequently their symbiotic competence. Furthermore, though different from plastid evolution, there is evidence of pseudogenization and genome erosion in the Azolla cyanobiont[5] but there is no evidence of gene exchange between cyanobiont and host genome.[32]

Insight on function of the HetR transcription factor in Trichormus azollae

Both NtcA and HetR transcription factors are induced in N-starved conditions and is triggered to action by 2-oxoglutarate. In the same conditions, a Hanks-type kinase (Pkn22) is induced that is capable of phosphorylation of specific residues of the HetR transcription factor for the differentiation of heterocysts from vegetative cells.[13] A bacterial two-hybrid system showed that HetR and Pkn22 interact with each other and mass spectrometry demonstrated that a conserved Ser-130 was phosphorylated in HetR upon Pkn22 interplay in all three oligomeric forms of HetR.[13] The Pkn22 expression is regulated by the NtcA transcription factor.

Up to 51 Hanks-type kinases are found in cyanobacteria, which suggest that there are other protein kinases that are able to phosphorylate key proteins such as HetR.[13] The Netphosbac 1.0 server which specializes in the prediction of prokaryotic phosphorylation sites, was employed for the identification of key residues that act as substrates to kinases. Using Netphosbac 1.0, six serine residues which are likely to be phosphorylated (Figure 6) were identified. Five were found in mid sequence, while one was found at the beginning of the *Trichormus azollae* HetR sequence (Figure 6). The six phosphorylation sites were found at 14, 121, 127, 166, 193 and 201 locations along the sequence of the *Trichormus azollae* HetR protein (Figure 6). However, Serine-130 was not predicted by the Netphosbac 1.0 portal, showcasing that prediction services have limitations in their functional assignment.

Five out of the six phosphorylation sites (Figures 6 and 7) were conserved between the HetR proteins of *Trichormus azollae* and *Nostoc punctiforme*, the former being a vertically transmitted cyanobiont and the latter a more promiscuous horizontally-transferred cyanobiont. Phosphorylation of the HetR protein in Nostoc PCC 7120 was shown to be effected by a Pkn22 kinase that is able to phosphorylate a highly conserved Serine-130 (TSLT**S**) that is conserved between heterocyst forming cyanobacteria.[13]

This five-residue motif (TSLT**S**) is conserved in subsection IV and V cyanobacteria suggesting a crucial sequence motif for induction of heterocyst formation.[13] In contrast, the subsection III cyanobacteria have a highly divergent motif (Figure 8) where a key serine residue (Serine-127) identified by Netphosbac 1.0

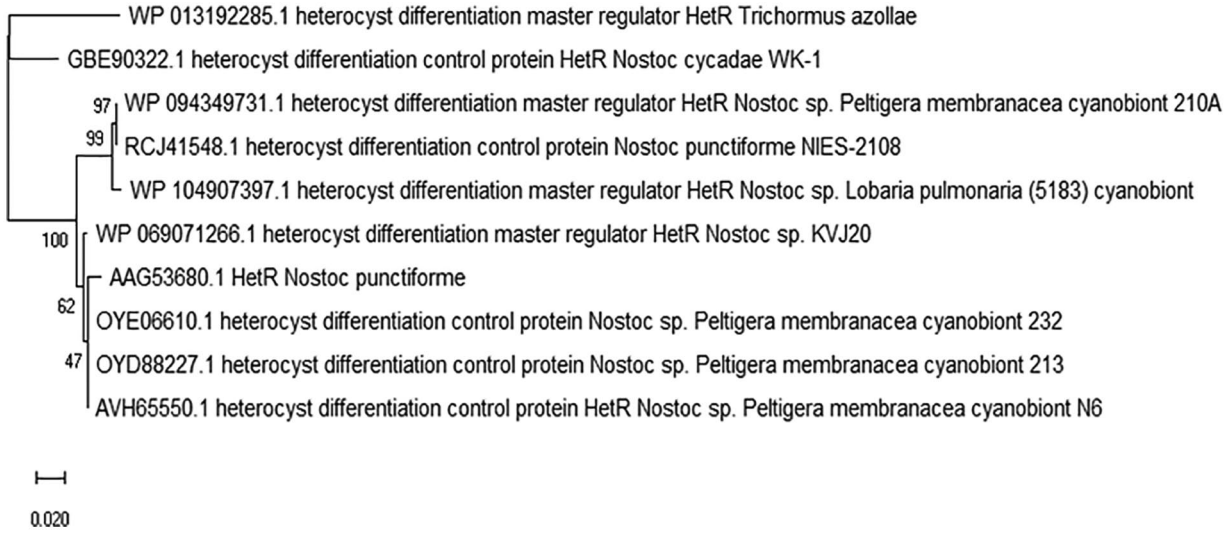


Figure 5. Phylogeny of HetR transcription factors from cyanobionts: The amino acid sequences of 10 cyanobionts were first aligned and the phylogenetic tree constructed using the Maximum Likelihood method, with bootstrap support from 500 replications.

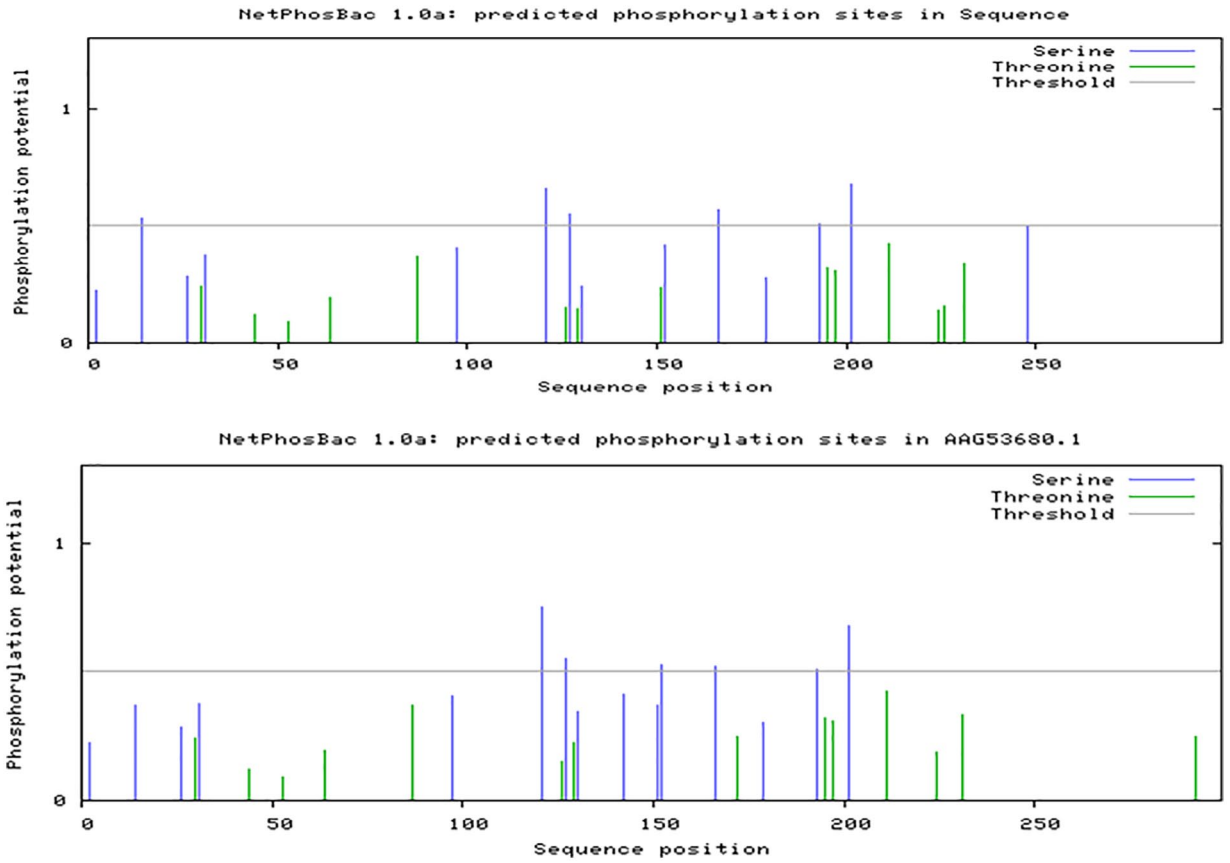


Figure 6. The prediction of serine and threonine phosphorylation sites for the HetR proteins from *Trichormus azollae* (top) and *Nostoc punctiforme* (bottom) using Netphosbac 1.0. Five out of the six serines that rise above the cutoff (horizontal line) are conserved between the two sequences. The X axis shows amino acid position and the Y axis demonstrates phosphorylation potential.

prediction, is transformed to an asparagine, hinting that this may be a significant mutation, for the absence of differentiation of heterocysts within this subsection. Asparagines, however, are unable to form phosphomimetic structures to be surrogates for serine phosphorylation (Figure 8). Strong heterocyst formation are conspicuous in phosphomimetic strains that have a constitutively active HetR protein.[13] Furthermore, Serine-130 of the HetR protein of subsection IV and V cyanobacteria is mutated to polar but neutral threonines, primary amino-group deficient prolines and hydrophobic alanines as well as valines, in subsection III cyanobacteria (Figure 8).

Serine is known to possess two disjointed codon types TCN (TCA, TCC, TCG, TCT) and AGY (AGT and AGC), which are farther apart than a single nucleotide substitution. Interestingly, both of Ser-127 and Ser-130 of the conserved 5 residue sequence in *Trichormus azollae*, are encoded by the same AGC codon (**agc[127]** ttg aca **agc [130]**) which means that the serine to asparagine substitution is a simple single-substitution based one. An induced serine→asparagine mutation of Ser-179 of the HetR protein abolished the protease function, showcasing that Ser-179 is perhaps the likely nucleophile that is central for the cleavage of the protein backbone[33] Serine to asparagine is the same mutation that occurs at Ser-127 in the T**S**LTS sequence of subsection III cyanobacteria (Figure 8) again symbolizing its importance as well as its ease of mutation.

Conclusion

The NtcA protein in *Trichmormus azollae* appears to be forming a dimer anchored by an intermolecular disulfide bond, which the other cyanobacteria appear to lack. The third cysteine I infer to be an important mutation. In the HetR proteins, there is a conserved patch of 5 residues which are conserved in all cyanobacteria capable of forming heterocysts, which is strongly mutated in subsection III cyanobacteria, which are capable of forming filaments but are unable to form specialized cells, with the exception of the genus *Trichodesmium* that forms diazocytes. This study advances the field in relation to (1) the selective phylogeny of cyanobionts using NtcA and HetR proteins and (2) structural and functional roles of NtcA and HetR proteins which could play a role in nitrogen metabolism; while presenting many more questions to be pursued empirically in the future.

Materials and Methods

Phylogenetic reconstructions

The non-redundant downloaded amino acid sequences (as FASTA files) from each query were first aligned with the ClustalW algorithm using MEGA version X (default parameters)[34] then converted to the MEGA sequence format, and phylogenetic reconstruction performed using the Neighbor Joining/Maximum l=Likelihood methods with support from 250, 500 or 1000 bootstrap replications. There was no assignment of outgroups.

```
# netphosbac-1.0a prediction results
#
# Sequence                 # x   Context     Score   Kinase    Answer
# ------------------------------------------------------------------
# Sequence                  2 S   ---MSNDID   0.221   main         .
# Sequence                 14 S   RLGPSAMDQ   0.534   main         Y
# Sequence                 26 S   YLAFSAMRT   0.282   main         .
# Sequence                 30 T   SAMRTSGHR   0.244   main         .
# Sequence                 31 S   AMRTSGHRH   0.375   main         .
# Sequence                 44 T   DAAATAAKC   0.121   main         .
# Sequence                 53 T   AIYMTYLEQ   0.090   main         .
# Sequence                 64 T   NLRMTGHLH   0.191   main         .
# Sequence                 87 T   RQALTEGKL   0.366   main         .
# Sequence                 97 S   KMLGSQEPR   0.403   main         .
# Sequence                121 S   HPGRSRVPG   0.657   main         Y
# Sequence                126 T   RVPGTSLTS   0.153   main         .
# Sequence                127 S   VPGTSLTSE   0.550   main         Y
# Sequence                129 T   GTSLTSEEK   0.144   main         .
# Sequence                130 S   TSLTSEEKK   0.242   main         .
# Sequence                151 T   AQLVTSFEF   0.236   main         .
# Sequence                152 S   QLVTSFEFL   0.419   main         .
# Sequence                166 S   LHKRSQEEL   0.570   main         Y
# Sequence                179 S   QMPLSEALA   0.278   main         .
# Sequence                193 S   RLLYSGTVT   0.510   main         Y
# Sequence                195 T   LYSGTVTRI   0.323   main         .
# Sequence                197 T   SGTVTRIDS   0.310   main         .
# Sequence                201 S   TRIDSPWGM   0.676   main         Y
# Sequence                211 T   FYALTRPFY   0.425   main         .
# Sequence                224 T   DQERTYTMV   0.142   main         .
# Sequence                226 T   ERTYTMVED   0.160   main         .
# Sequence                231 T   MVEDTARYF   0.338   main         .
# Sequence                248 S   RKANSMRAV   0.494   main         .
#
    MSNDIDLIKRLGPSAMDQIMLYLAFSAMRTSGHRHGAFLDAAATAAKCAI   #     50
    YMTYLEQGQNLRMTGHLHHLEPKRVKIIVEEVRQALTEGKLLKMLGSQEP   #    100
    RYLIQLPYLWMEKYPWHPGRSRVPGTSLTSEEKKQIERKLPKNLPDAQLV   #    150
    TSFEFLELIEFLHKRSQEELPPHHQMPLSEALAEHIKRRLLYSGTVTRID   #    200
    SPWGMPFYALTRPFYAPADDQERTYTMVEDTARYFRMMKDWAERKANSMR   #    250
    AVEELDIPIEQMQQAMEELDEIIRAWADKYHQDGGMPMVLQMVFANQDQ    #    300
%1  ............S.....................................    #     50
%1  ................................................    #    100
%1  ..................S.....S.......................    #    150
%1  ..............S.........................S.......    #    200
%1  S................................................    #    250
%1  ................................................
```

Figure 7. The predicted serines that were shown to be putative substrates for phosphorylation by an upstream kinase, shown using the Netphosbac 1.0 server. The positive ones are marked Y and coded in yellow color.

Secondary structure prediction

The secondary structure prediction service PSIPRED 4.0 (http://bioinf.cs.ucl.ac.uk/psipred/) was used to showcase the helixes, beta strands, and coils.[35]

Homology modeling

Homology modeling was performed using the default parameters of the SWISS-MODEL server (https://swissmodel.expasy.org/).[36,37]

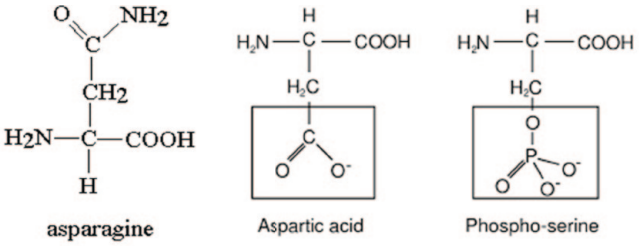
Phosphorylation site prediction

The web address (http://www.cbs.dtu.dk/services/NetPhosBac/) hosting the Netphosbac 1.0[38] was used for the identification of the putative phosphorylation sites.

Prediction of disulfide bonds

The selected sequence was searched against the DiANNA server (http://clavius.bc.edu/~clotelab/DiANNA/) for the identification of likely disulfide bond pairs.[29]

A



B

2. AAL05045.1 heterocyst differentiation protein Trichodesmium erythraeum IMS101	S R L P G N N L T T E E K R
3. AAB81942.1 heterocyst differentiation protein partial Trichodesmium tenue	S R L P G N N L T T E E R K
4. AAB81941.1 heterocyst differentiation protein partial Trichodesmium contortum	S R L P G N N L P T E E R K
5. AAQ06415.1 heterocyst differentiation protein partial Trichodesmium aureum	P R V P G S N L T T E E K R
6. CBN56579.1 Heterocyst differentiation control protein Oscillatoria sp. PCC 6506	A R I P G N N L T A E E K R
7. WP_015148088.1 heterocyst differentiation master regulator HetR Oscillatoria acuminata	P R I P G N N L T S E E K R
8. WP_158307959.1 heterocyst differentiation master regulator HetR Microcoleus vaginatus	P R I P G N N L T V D E K K
9. WP_006618511.1 heterocyst differentiation master regulator HetR Arthrospira platensis	A R V P G S S L T Q E E K R
10. WP_096831969.1 heterocyst differentiation master regulator HetR Tychonema bourrellyi	P R I P G N S L T A D E K K
11. AAQ06413.1 heterocyst differentiation protein partial Trichodesmium pelagicum	P R V P G G N L T T E E K G
12. NES94832.1 heterocyst differentiation master regulator HetR Desertifilum sp. SIO1I2	S R I P G T S L T S E E K R
13. WP_017318187.1 heterocyst differentiation master regulator HetR Mastigocladopsis repens	A R V P G T N L T P E E K R
14. WP_041268935.1 heterocyst differentiation master regulator HetR Geitlerinema sp. PCC 7407	P R I P G T S L T S E E K R
15. AAD43512.1 HetR partial Nostoc sp. GP 9401	S R V P G T S L T S E E K R
16. WP_094349731.1 heterocyst differentiation master regulator HetR Nostoc sp. Peltigera membranace	S R V P G T N L T S E E K R
17. AAD43509.1 HetR partial Calothrix sp. PCC 7507	S R I P G T S L T T E E K K
18. ADM83442.1 heterocyst differentiation partial Dolichospermum flos-aquae PMC207.03	S R V P G T S L T S E E K K
19. WP_069071266.1 heterocyst differentiation master regulator HetR Nostoc sp. KVJ20	S R V P G S S L T S E E K R
20. BAE80661.1 heterocyst differentiation protein partial Desmonostoc sp. PCC 7906	S R V P G T S L T S E E K R
21. BAE80658.1 heterocyst differentiation protein partial Fischerella muscicola PCC 7414	S R I P G T S L T S E E K R
22. BAE80698.1 heterocyst differentiation protein partial Fischerella thermalis PCC 7521	S R I P G T S L T S E E K R
23. AUB37256.1 peptidase S48 HetR Nostoc flagelliforme CCNUN1	S R V P G T S L T S E E K R
24. ADM83445.1 heterocyst differentiation partial Planktothrix sp. PMC253.05	S R I S G S S L T S E E K R
25. BAE80657.1 heterocyst differentiation protein partial Fischerella muscicola SAG 1427-1 PCC 73103	S R I P G T S L T S E E K R
26. AAM89674.1 HetR partial Aphanizomenon flos-aquae TR183	S R V P G T S L T S E E K K
27. AAG53966.1 HetR partial Nodularia harveyana	S R V P G T N L T S E E K R
28. BAE80663.1 heterocyst differentiation protein partial Tolypothrix sp. CCMP1185	S R I P G T S L T S E E K R
29. BAE80659.1 heterocyst differentiation protein partial Fischerella muscicola UTEX 1829	S R I P G T S L T S E E K R
30. ERT05802.1 heterocyst differentiation control protein Lyngbya aestuarii BL J	P R I P G N S L T P E E K R
31. WP_049560785.1 heterocyst differentiation master regulator HetR Limnoraphis robusta	P R I P G N G L T P E E K R

C

Leptolyngbya sp. PCC 73110	- NNLTT
Trichodesmium erythraeum IMS101	- NNLTT
Trichodesmium tenue	- NNLTT
Trichodesmium contortum	- NNLPT
Trichodesmium aureum	- SNLTT
Trichodesmium pelagicum	-GNLTT
Oscillatoria sp. PCC 6506	-NNLTA
Oscillatoria acuminate	-NNLTS
Microcoleus vaginatus	-NNLTV
Tychonema bourrellyi	-NSLTA
Leptolyngbya ohadii	-GNLSS
Limnoraphis robusta	-NGLTP
Lyngbya aestuarii BL J	-NSLTP
Trichormus azollae	-TSLTS

Figure 8. The structures of the amino acids asparagine, phosphomimetic aspartic acid, and phosphoserine. (B) Sequence alignment of HetR proteins showing the conserved **TSLTS** motif, which is strongly divergent in subsection III non-heterocystous cyanobacteria. (C) The highly mutated five amino acid motif in subsection III cyanobacteria shown against that of subsection IV *Trichormus azollae*.

Multiple sequence alignments

The non-redundant downloaded amino acid sequences (as FASTA files) were employed for sequence alignment using the ClustalW algorithm using the MEGA X software.[34]

Author Contributions

Dr. Dilantha Gunwardana is responsible for the conceptualization, methodology, investigation and writing of the manuscript. Michelle Alexander illustrated Figure 4(D).

ORCID iD

Dilantha Gunawardana https://orcid.org/0000-0002-5086-0215

REFERENCES

1. Zheng WW, Bergman B, Chen B, Zheng S, Xiang G, Rasmussen U. Cellular responses in the cyanobacterial symbiont during its vertical transfer between plant generations in the Azolla microphylla symbiosis. *New Phytol*. 2009;181:53-61.
2. Kollah BB, Patra AK, Mohanty SR. Aquatic microphylla Azolla: a perspective paradigm for sustainable agriculture, environment and global climate change. *Environ Sci Pollut Res Int*. 2016;23:4358-4369.
3. Wagner GM. Azolla: a review of its biology and utilization. *Bot Rev*. 1997;63:1-26.
4. Pereira ALAL, Vasconcelos V. Classification and phylogeny of the cyanobiont Anabaena azollae Strasburger: an answered question? *Int J Syst Evol Microbiol*. 2014;64:1830-1840.
5. Ran L, Larsson J, Vigil-Stenman T, et al. Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PLoS ONE*. 2010;5:e11486.
6. Ekman MM, Tollback P, Bergman B. Proteomic analysis of the cyanobacterium of the Azolla symbiosis: identity, adaptation, and NifH modification. *J Exp Bot*. 2008;59:1023-1034.
7. Gunawardana D. An exploration of common greenhouse gas emissions by the cyanobiont of the azolla-nostoc symbiosis and clues as to nod factors in cyanobacteria. *Plants*. 2019;8:587.
8. Sood A, Uniyal PLPL, Prasanna R, Ahluwalia AS. Phytoremediation potential of aquatic macrophyte, Azolla. *Ambio*. 2012;41:122-137.
9. Ehira S, Ohmori M. NrrA directly regulates expression of hetR during heterocyst differentiation in the cyanobacterium Anabaena sp. strain PCC 7120. *J Bacteriol*. 2006;188:8520-8525.
10. Ehira S, Ohmori M. NrrA, a nitrogen-responsive response regulator facilitates heterocyst development in the cyanobacterium Anabaena sp. strain PCC 7120. *Mol Microbiol*. 2006;59:1692-1703.
11. Llacer JL, Espinosa J, Castells MA, Contreras A, Forchhammer K, Rubio V. Structural basis for the regulation of NtcA-dependent transcription by proteins PipX and PII. *Proc Natl Acad Sci USA*. 2010;107:15397-15402.
12. Zhao MX, Jiang YL, He YX, et al. Structural basis for the allosteric control of the global transcription factor NtcA by the nitrogen starvation signal 2-oxoglutarate. *Proc Natl Acad Sci USA*. 2010;107:12487-12492.
13. Roumezi B, Xu X, Risoul V, Fan Y, Lebrun R, Latifi A. The Pkn22 Kinase of Nostoc PCC 7120 is required for cell differentiation via the Phosphorylation of HetR on a residue highly conserved in genomes of heterocyst-forming cyanobacteria. *Front Microbiol*. 2019;10:3140.
14. Wang C-M, Ekman M, Bergman B. Expression of cyanobacterial genes involved in heterocyst differentiation and dinitrogen fixation along a plant symbiosis development profile. *Mol Plant Microbe Interact*. 2004;17:436-443.
15. de Mendoza A, Sebe-Pedros A, Sestak MS, et al. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc Natl Acad Sci U S A*. 2013;110:E4858-4866. doi:10.1073/pnas.1311818110.
16. Picossi S, Flores E, Herrero A. ChIP analysis unravels an exceptionally wide distribution of DNA binding sites for the NtcA transcription factor in a heterocyst-forming cyanobacterium. *BMC Genomics*. 2014;15:22.
17. Videau P, Rivers OS, Tom SK, et al. The hetZ gene indirectly regulates heterocyst development at the level of pattern formation in Anabaena sp. strain PCC 7120 [published online ahead of print April 20, 2018]. *Mol Microbiol*. doi:10.1111/mmi.13974.
18. Zhang WW, Du Y, Khudyakov I, et al. A gene cluster that regulates both heterocyst differentiation and pattern formation in Anabaena sp. strain PCC 7120. *Mol Microbiol*. 2007;66:1429-1443.
19. Atugoda DRAMTR, Mandakini LLU, Bandara NJGJ, Gunawardana D. How a taxonomically-ambiguous cyanobiont and vanadate assist in the phytoremediation of cadmium by Azolla Pinnata: implications for CKDU. *Environ Pollut*. 2018;7:53-65.
20. Warshan D, Liaimer A, Pederson E, et al. Genomic changes associated with the evolutionary transitions of Nostoc to a plant symbiont. *Mole Biol Evol*. 2018;35:1160-1175. doi:10.1093/molbev/msy029.
21. Bell-Doyon P, Laroche J, Saltonstall K, Villarreal Aguilar JC. Specialized bacteriome uncovered the coralloid roots of the epiphytic gymnosperm Zamia pseudoparasitica. *Environ DNA*. 2020;2:418-428. doi:10.1002/edn3.66.
22. Wisen S, Bergman B, Mannervik B. Mutagenesis of the cysteine residues in the transcription factor NtcA from Anabaena PCC 7120 and its effects on DNA binding in vitro. *Biochim Biophys Acta*. 2004;1679:156-163.
23. Sandmann G, Mautz J, Breitenbach J. Control of light-dependent keto carotenoid biosynthesis in Nostoc 7120 by the transcription factor NtcA. *Z Naturforsch C J Biosci*. 2016;71:303-311.
24. Poole LB. The basics of thiols and cysteines in redox biology and chemistry. *Free Radic Biol Med*. 2015;80:148-157.
25. Xia X, Xie Z. Protein structure, neighbor effect, and a new index of amino acid dissimilarities. *Mol Biol Evol*. 2002;19:58-67.
26. Xia X. DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. *Mol Biol Evol*. 2013;30:1720-1728.
27. Grantham RR. Amino acid difference formula to help explain protein evolution. *Science* 1974;185:862-864.
28. Miyata T, Miyazawa S, Yasunaga T. Two types of amino acid substitutions in protein evolution. *J Mol Evol* 1979;12:219-236.
29. Ferre F, Clote P. DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Res*. 2005;33:W230-W232.
30. Vamosi G, Baudendistel N, von der Lieth CW, et al. Conformation of the c-Fos/c-Jun complex in vivo: a combined FRET, FCCS, and MD-modeling study. *Biophys J*. 2008;94:2859-2868.
31. Szaloki N, Krieger JW, Komaromi I, Toth K, Vamosi G. Evidence for homodimerization of the c-Fos transcription factor in live cells revealed by fluorescence microscopy and computer modeling. *Mol Cell Biol*. 2015;35:3785-3798.
32. Li F, Brouwer P, Carretero-Paulet, et al. Fern genomes elucidate land plant evolution cyanobacterial symbioses. *Nature Plants* 2018;4:460-472. doi:10.1038/s41477-018-0188-8.
33. Zhou R, Wei X, Jiang N, et al. Evidence that HetR protein is an unusual serine-type protease. *Proc Natl Acad Sci USA*. 1998;95:4959-4963.
34. Sohpal VK, Dey A, Singh A. MEGA biocentric software for sequence and phylogenetic analysis: a review. *Int J Bioinform Res Appl*. 2010;6:230-240.
35. McGuffin LJLJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics*. 2000;16:404-405.
36. Bienert S, Waterhouse A, de Beer TA, et al. The SWISS-MODEL repository-new features and functionality. *Nucleic Acids Res*. 2017;45:D313-D319.
37. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. 2018;46:W296-W303.
38. Miller MLML, Soufi B, Jers C, Blom N, Macek B, Mijakovic I. NetPhosBac—a predictor for Ser/Thr phosphorylation sites in bacterial proteins. *Proteomics*. 2009;9:116-125.