Identification of hypothetical proteins with putative arsenate reductase properties in cyanobacteria by bioinformatics approach

PV Parvati Sai Arun*

Centre for Bioinformatics, CR Rao Advanced Institute of Mathematics, Statistics and Computer science, Hyderabad, Telangana, India.

Received: May 30, 2017; accepted: July 25, 2017.

This study focuses on the identification of proteins, which are annotated as hypothetical proteins, but possess putative arsenate reductase properties among cyanobacteria, by using bioinformatics approach. In the present work, we have chosen the protein sequence of the gene *all0195*, which was annotated as hypothetical protein and later identified as arsenate reductase in *Anabaena sp.* PCC 7120. For this selected reference protein, we searched for conserved orthologs among other 74 sequenced cyanobacteria using the bidirectional best hits method. A total of seven hypothetical proteins were identified as bidirectional best hits for the protein All0195 of *Anabaena sp.* PCC 7120 across the 74 sequenced cyanobacterial species. These protein sequences of the predicted bidirectional hits were further in-depth analyzed using different bioinformatics tools. From the in-depth bioinformatics analysis, it was observed that the hypothetical proteins, which were identified by using bioinformatics approach, were found to have the properties of arsenate reductase proteins and were very similar to the protein All0195 of *Anabaena*.

Keywords: Arsenate reductase; hypothetical proteins; sequenced cyanobacterial species; bidirectional best hits; bioinformatics approach.

***Corresponding author:** PV Parvati Sai Arun, Center for Bioinformatics, CR Rao Advanced Institute of Mathematics, Statistics and Computer science, University of Hyderabad Campus, Hyderabad, Telangana 500046, India. Email: arun.uoh@gmail.com.

Introduction

Cyanobacteria are photosynthetic organisms believed to be the oldest forms of life existing on earth. They are widely distributed in different environments such as aquatic, hot springs, deserts, and polar environments [1]. They are considered as the globally important primary producers and also as the progenitors of plant chloroplasts [2-4]. They possess vital metabolic pathways and survival mechanisms, and hence became important model systems [5]. As cyanobacteria is present in a wide variety of ecological niches, they are naturally encounter to different kinds of metals present in their niche. Arsenic is one of such metal, which is highly toxic and present abundantly in different ecological niches [6]. Arsenic is a group V metalloid. When present in higher concentrations, it will lead to the abiotic stress in many plants, cyanobacteria, and also in other forms of life [7-9]. To counteract the toxic effects of the arsenic, cyanobacteria possess arsenic resistant genes organized in the form of Operons in the order of *arsRBDAC* on their chromosomes or in the plasmid [10]. The gene *arsR* encodes for the repressor protein, whereas *arsB* encodes for the membrane arsenite permease pump. The key enzyme involved in the detoxification reaction of arsenate to arsenite, the arsenate reductase, is encoded by *arsC* gene [8, 11]. It was reported that there exists more than one gene encoding for arsenate reductase in cyanobacteria. For example, in the

cyanobacterium *Anabaena sp.* PCC 7120, according to the annotation of its genome, the proteins encoded by the genes *alr1105* and *alr2520* were reported as they belong to arsenate reductase family [8]. Later, a "hypothetical protein" encoded by the gene *all0915* was reported to have arsenate reductase properties in the same organism. This evidence confirms that there exists more than one gene in the genome of cyanobacteria, such as *Anabaena*, encoding for arsenate reductase like proteins, but were annotated as "hypothetical proteins". For the identification of such hypothetical proteins which have putative arsenate reductase properties across the sequenced cyanobacterial proteomes, we used different bioinformatics techniques and predicted putative arsenate reductase like proteins, which are similar to *all0195* (coding for arsenate reductase) in *Anabaena*. In the process of identification of these new putative arsenate reductase proteins across sequenced cyanobacterial proteomes, we have adopted the following strategy 1) Identification of bidirectional best hits of the protein All0195; 2) Prediction of Physical-chemical properties of All0195 and its predicted bidirectional best hits; 3) Performing the primary sequence alignment followed by secondary structure prediction; 4) Prediction of tertiary structure and validation.

Material and methods

In this study a total of 74 cyanobacterial proteomes were considered. The *.faa files of all these 74 cyanobacterial proteomes were downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/). Local protein database of all these 74 proteomes were constructed using "makeblastdb" module of BLAST+ package. The protein All0195 of *Anabaena* which was demonstrated to have the arsenate reductase properties was considered as the reference protein for which bidirectional hits were predicted. For the prediction of bidirectional best hits, forward and reverse BLASTP was performed between protein sequence of All0195 and other cyanobacterial proteomes (the local database) [12]. After the prediction of bidirectional best hits, the protein sequences of the predicted bidirectional best hits along with All0195 were retrieved from the local protein database using "blastdbcmd" module of BLAST+ package. The bidirectional best hits for the protein All0195 of *Anabaena* which have their annotation as "hypothetical proteins" were considered for further analysis. The resulting "hypothetical protein" sequences, including the protein sequence of All0195 (here after called as primary seed), were given as input to PEPSTATS of EMBOSS package for the prediction of physical-chemical properties of the proteins [13]. Aliphatic index and GRAVY were computed using in house Perl program using the standard mathematical equations described earlier [14, 15]. The T-coffee server was used for performing the multiple sequence alignment of the primary seed [16]. Protein domain analysis for this primary seed was done using SMART database including Pfam searches [17, 18]. Secondary structure elements were predicted using CFSSP server [19, 20]. To perform homology modelling, the proteins present in the primary seed were split into individual protein sequences, and BLASTP was performed against a PDB database, for the identification of suitable templates. The PDB files of the suitable templates were downloaded from RCSB database (http://www.rcsb.org/pdb/home/home.do).
The individual proteins obtained from the primary seed, along with its template, were given as input for Modeller version 9.10 [21]. Structure validation for the generated homology models was performed using the RAMPAGE server [22]. The visualization of the models and super imposition of the models with their templates were performed with Pymol (The PyMOL Molecular Graphics System (2002) by W. L. Delano).

Results

For the prediction of conserved orthologs

Table 1. Predicted bidirectional best hits of All0195 protein of *Anabaena sp*. PCC 7120.

No.	ORF ID	Gene name annotated	Cyanobacterium	Protein function annotated
1	*sync_2016*	-	*Synechococcus* CC9311	Arsenate reductase
2	*p9515_05761*	*arsC*	*Prochlorococcus marinus* MIT 9515	Arsenate reductase
3	*a9601_05691*	*arsC*	*Prochlorococcus marinus* AS9601	Arsenate reductase
4	*p9303_07481*	*arsC*	*Prochlorococcus marinus* MIT 9303	Arsenate reductase
5	*natl1_05701*	*arsC*	*Prochlorococcus marinus* NATL1A	Arsenate reductase
6	*p9301_05391*	*arsC*	*Prochlorococcus marinus* MIT 9301	Arsenate reductase
7	*synwh7803_0626*	-	*Synechococcus* WH 7803	Arsenate reductase C
8	*p9215_05941*	*arsC*	*Prochlorococcus marinus* MIT 9215	Putative arsenate reductase
9	***am1_5153***	-	*Acaryochloris marina* MBIC11017	Hypothetical protein
10	*p9211_05131*	*arsC*	*Prochlorococcus marinus* MIT 9211	Arsenate reductase
11	***synpcc7002_a0009***	-	*Synechococcus* PCC 7002	Hypothetical protein
12	*all0195*	-	*Anabaena* sp. PCC 7120	Arsenate reductase
13	*npun_f2949*	-	*Nostoc punctiforme* PCC 73102	Arsenate reductase
14	*cyan7425_0736*	-	*Cyanothece* PCC 7425	Arsenate reductase
15	***cyan7822_0675***	-	*Cyanothece* PCC 7822	Hypothetical protein
16	*pro_0511*	*arsC*	*Prochlorococcus marinus* CCMP1375	Arsenate reductase related protein
17	*pmm0512*	-	*Prochlorococcus marinus pastoris* CCMP1986	Arsenate reductase
18	*pmt1256*	-	*Prochlorococcus marinus* MIT 9313	Arsenate reductase
19	*synw1767*	-	*Synechococcus* WH 8102	Arsenate reductase
20	*cyagr_0067**	-	*Cyanobium gracile* PCC 6307	Spx/MgsR family transcriptional regulator
21	*nos7107_4768*	-	*Nostoc* PCC 7107	ArsC family protein
22	*cal7507_1001*	-	*Calothrix* PCC 7507	ArsC family protein
23	*lepto7376_1744*	-	*Leptolyngbya* PCC 7376	ArsC family protein
24	*riv7116_6657**	-	*Rivularia* PCC 7116	Spx/MgsR family transcriptional regulator
25	*pse7367_0067*	-	*Pseudanabaena* PCC 7367	Arsenate reductase
26	*gei7407_2028*	-	*Geitlerinema* PCC 7407	Arsenate reductase
27	*cal6303_4586*	-	*Calothrix* PCC 6303	ArsC family protein
28	*mic7113_5134**	-	*Microcoleus* PCC 7113	Spx/MgsR family transcriptional regulator
29	*cyan10605_2070*	-	*Cyanobacterium aponinum* PCC 10605	ArsC family protein
30	*cyast_2291*	-	*Cyanobacterium stanieri* PCC 7202	Arsenate reductase
31	*cha6605_5314**	-	*Chamaesiphon minutus* PCC 6605	Transcriptional regulator, Spx/MgsR family
32	*anacy_1198*	-	*Anabaena cylindrica* PCC 7122	ArsC family protein
33	***nies39_a00120***	-	*Arthrospira platensis* NIES 39	Putative uncharacterized protein, fragment
34	***pmn2a_1845***	-	*Prochlorococcus marinus* NATL2A	Hypothetical protein
35	***ava_2687***	-	*Anabaena variabilis* ATCC 29413	Hypothetical protein
36	***syncc9902_1661***	-	*Synechococcus* CC9902	Hypothetical protein
37	***syncc9605_0697***	-	*Synechococcus* CC9605	Hypothetical protein
38	***pmt9312_0513***	-	*Prochlorococcus marinus* MIT 9312	Hypothetical protein

Notes: The ORF IDs in bold letters are the proteins which are predicted as bidirectional best hits of All0195 of *Anabaena* but were annotated as Hypothetical proteins. The ORF IDs with * are the proteins predicted as bidirectional best hits of All0195 but annotated to have different function.

Table 2. Conserved Domain analysis of the protein All0195 and its bidirectional best hits with SMART database.

Protein ID	Function	Organism and adaptation	Predicted Domain	Position of the Domain	Amino acids sequence present in the domain
All0195	**Arsenate reductase**	**Nostoc PCC 7120 (Soil)**	**ArsC**	**6-112**	**YGIPNCGTCKKTFNWLQAHKV DYEFINTKENPPTREHIQNWVK SLSSTPMRNTSGQSYRALGEEK KNWTDEQWIEEFAKDAMLLKR PLFVKDGIAVAVGFRDEKIIR**
Am1_5153	Hypothetical Protein	*Acaryochloris marina* MBIC11017 (Marine)	ArsC	6-113	YGIPTCGTCKKALKWLQENQLE FEFINTKEEPPSIQQISAWVDTF GSKPMRNTSGGAYRALGEQKK TWSEDQWIAAFAEDAMLLKRP LILKDGAPVLVGFRASDEVLK
Ava_2687	Hypothetical protein	*Anabaena variabilis* ATCC 29413 (Fresh water)	ArsC	6-112	YGIPNCGTCKKAFNWLQAHKV DYEFINTKENPPTRENIQNWVK SLGSTPMRNTSGQSYRALGEEK KNWTDEQWIEEFAKDAMLLKR PLFVKDGIAVAVGFRDEKVIQ
Nies39_a00120*	Putative uncharacterized protein	*Arthrospira platensis* NIES-39 (Salt water lake)	Not Identified	--	--
Cyan7822_0675*	Hypothetical protein	*Cyanothece sp*. PCC 7822 (Soil)	Not identified	--	--
Pmt9312_0513	Hypothetical protein	*Prochlorococcus marinus* str. MIT 9312 (Marine)	ArsC	7-116	YSYLKCSTCRKAAKWLDKKDFEY QLIDIVKEPPLLDYLNLALEQYSP DKKRIFNTRGKAFKSINLDIYSLS KEEIIQLLLSDGKLIKRPFLVYEEK KVILGFNEIEYAEQ
Pmn2a_1845	Hypothetical protein	*Prochlorococcus marinus* str. NATL2A (Marine)	ArsC	4-113	FSYSSCSTCRRAIKWLKYNDIPFE LIDLLKSPPSKEMLISASELYGDR KYLLNTSGVVYRSMGSDAVKK MSDNDLFEQLILEPRLIKRPFLYK SSKCFLVGFKEEKWAEK
Syncc9605_0697	Hypothetical protein	*Synechococcus sp*. CC9605 (Marine)	ArsC	8-117	YSYNRCSTCRKALAWLTERGIAH EVHDITLTPPSKDMLVAAHQSL GDRKLLFNTSGQSYRAMGAAA VKALSDDEALEALAADGKLIKRP FVEVNSSTYLTGFKPDLWESS
Syncc9902_1661	Hypothetical protein	*Synechococcus sp*. CC9902 (Marine)	ArsC	8-116	YSYNRCSTCRKALAWLTDQGIA HDVHDIVENPPSRNDLDAAFAF LGDRKLLFNTSGQSYRALGSAVV KAMSDSEALAALAKDGKLIKRPF VKRSDGSFLVGFKPEVWAS
Synpcc7002_a0009	Hypothetical protein	*Synechococcus sp*. PCC 7002 (Sediment)	ArsC	6-112	YGIPTCNTCKKALKWLETAGISY EFINTKEQPPTRQAIAQWVSDL GSKPMRNTSGQSYRALGEEKKT WDDNQWIEAFSQDAMLLKRPL FVRDNKAVLVGFRASETEL

Notes: For the proteins marked with * no domain was identified by SMART database including the option of Pfam search. The protein in bold is the reference protein All0195 of *Anabaena* PCC 7120.

between related species, the bidirectional best hit method is widely used in many studies [23-25]. In this study, the bidirectional best hits for the protein sequence of All0195 of *Anabaena sp*. were predicted among the other 74 cyanobacterial species. A total of 38 bidirectional best hits were predicted among 74 cyanobacterial species for the protein All0195 (table 1). Of these 38 bidirectional best hits, there are 9 proteins annotated as "hypothetical proteins" and 29 proteins annotated as arsenate reductase (table 1). The multiple sequence alignment of these 9 hypothetical proteins along with the All0195 protein sequence (the primary seed) reveals that most of the protein sequences of the bidirectional hits were conserved except for the protein sequences of Nies39_A00120 and Cyan7822_0675 of *Arthrospira platensis* NIES 39 and *Cyanothece* PCC 7822 respectively (data not shown).

Protein domain analysis

The protein sequences of the primary seed were given as input individually for SMART database. By default, SMART searches for the presence of conserved protein domains in the submitted protein sequence using Hidden Markov models (HMMER). In our analysis, we used both HMMER search along with Pfam search, which is present as an additional search parameter in SMART. Upon submitting the primary seed to SMART, except the proteins encoded by the genes *nies39_a00120*, *cyan7822_0675*, the rest of the genes which are annotated as hypothetical proteins were found to have ArsC domain, which is found in arsenate reductase present in cyanobacteria [26], in their primary protein sequence (table 2). As the proteins Nies39_A00120 and Cyan7822_0675 have less conservation of amino acids as observed in multiple sequence alignment and absence of the ArsC domain in their sequence, these two proteins are removed from further analysis. The remaining seven hypothetical proteins along with All0195 were considered for further analysis (hereafter referred as secondary seed). Multiple sequence alignment for the proteins present in the secondary seed was performed for one more time, where the results revealed good conservation among the amino acids of the proteins present in the secondary seed (figure 1).

Prediction of Physical-chemical properties

The secondary seed was given as input to the PEPSTATS tool of Emboss package. PEPSTATS predicted molecular weight, total number of residues, average residue weight, charge, isoelectric point, A280 molar extinction coefficients, and other properties. The Aliphatic Index, GRAVY value was predicted from the output generated by PEPSTATS. From the prediction of protein properties, it was revealed that the molecular weight of All0195 (reference protein) is 13.6 kDa. The observed molecular weights of the predicted bidirectional best hits of All0195 ranges from 13 kDa to 14 kDa. The total number of residues was ranging from 115 to 120 amino acids. The computed pI ranges from 6.3 in the case of Am1_5153 of *Acaryochloris marina* MBIC 11017 to 9.5 in the case of Pmn2a_1845 of *Prochlorococcus marinus* NATL2A. From the prediction of pI, it is clear that, out of seven hypothetical proteins, it can be assumed that the protein Am1_5153 of *Acaryochloris marina* MBIC11017 precipitates in acidic buffers since its pI is below 7, whereas rest of six precipitates in basic buffers since their pI is above 7 (data not shown). From the amino acids composition data, it was observed that, in these seven hypothetical proteins, leucine was found to be more predominant than the other amino acids followed by lysine and so on (data not shown). The aliphatic index indicates the relative volume occupied by aliphatic side chains [14]. The aliphatic index for the proteins present in the secondary seed ranges from 70 to 104. The GRAVY value was ranging from -0.19 to -0.65 indicating that the proteins better interact with water.

Prediction of secondary structure

We used CFSSP server for the prediction of secondary structure elements for the proteins in secondary seed. CFSSP server predicts helices, sheets and coils in the given protein sequence along with their percentage of amino acids

Figure 1. Multiple sequence alignment of the seven Hypothetical proteins and All0195. The color scale from Blue to Pink show the conservation of the amino acids.

Table 3. Secondary structure analysis of All0195 and its bidirectional best hits.

Protein ID	Function	Organism	Total Number of residues	Total residues in helix	Total residues in sheet	Total residues in coil
All0195	**Arsenate reductase**	***Anabaena* PCC 7120**	**117**	**89 (76.1%)**	**39 (33.3%)**	**17 (14.5%)**
Am1_5153	Hypothetical Protein	*Acaryochloris marina* MBIC11017	118	95 (80.5%)	45 (38.1%)	19 (16.1%)
Ava_2687	Hypothetical protein	*Anabaena variabilis* ATCC 29413	117	88 (75.2%)	35 (29.2%)	17 (14.5%)
Pmt9312_0513	Hypothetical protein	*Prochlorococcus marinus* str. MIT 9312	118	102 (86.4%)	59 (50%)	18 (15.3%)
Pmn2a_1845	Hypothetical protein	*Prochlorococcus marinus* str. NATL2A	115	77 (67%)	44 (38.3%)	15 (13%)
Syncc9605_0697	Hypothetical protein	*Synechococcus* sp. CC9605	120	86 (71.7%)	35 (29.2%)	20 (16.7%)
Syncc9902_1661	Hypothetical protein	*Synechococcus* sp. CC9902	120	88 (73.3%)	26 (21.7%)	21 (17.5%)
Synpcc7002_a0009	Hypothetical protein	*Synechococcus* sp. PCC 7002	119	86 (72.3%)	45 (37.8%)	19 (16%)

Note: The protein in bold is the target protein All0195 of *Anabaena*.

Table 4. Identified templates, their source and percentage similarity for the individual proteins present in the secondary seed for homology modeling.

Individual primary seed protein	Organism	Identified template	Percentage similarity	Template source	RMSD Value
All0195	***Anabaena* PCC 7120**	**3FZ4**	**76.6**	***Streptococcus mutans* Ua159**	--
AM1_5153	*Acaryochloris marina* MBIC11017	3FZ4	72.8	*Streptococcus mutans* Ua159	0.275
Ava_2687	*Anabaena variabilis ATCC* 29413	3FZ4	78.2	*Streptococcus mutans* Ua159	0.217
Pmt9312_0513	*Prochlorococcus marinus* MIT 9312	3FZ4	85.9	*Streptococcus mutans* Ua159	0.219
Pmn2a_1845	*Prochlorococcus marinus* NATL2A	3FZ4	59.3	*Streptococcus mutans* Ua159	0.266
Syncc9605_0697	*Synechococcus* CC9605	2M46	74.3	*Staphylococcus aureus* subsp. aureus COL	0.512
Syncc9902_1661	*Synechococcus* CC9902	3GKX	77.4	*Bacteroides fragilis*	0.306
Synpcc7002_A0009	*Synechococcus* PCC 7002	3FZ4	73.9	*Streptococcus mutans* Ua159	0.244

Notes: The protein in bold represent the target protein All0195 from *Anabaena*. The RMSD value of All0195 is not calculated since it is considered as reference protein.

Table 5. Ramachandran plot analysis of the homology models generated using Modeller.

Individual primary seed protein	Percentage of amino acids in most allowed regions	Percentage of amino acids in allowed regions	Percentage of amino acids in outlier region
All0195 (*Anabaeba* PCC 7120)	**98.3**	**1.7**	**0**
Am1_5153 (*Acaryochloris marina* MBIC11017)	94.8	2.6	2.6
Ava_2687 (*Anabaena variabilis ATCC* 29413)	94.8	3.5	1.7
Pmt9312_0513 (*Prochlorococcus marinus* MIT 9312)	95.7	1.7	2.6
Pmn2a_1845 (*Prochlorococcus marinus* NATL2A)	98.2	1.8	0
Syncc9605_0697 (*Synechococcus* CC9605)	94.9	3.4	1.7
Syncc9902_1661 (*Synechococcus* CC9902)	94.9	2.5	2.5
Synpcc7002_A0009 (*Synechococcus* PCC 7002)	97.4	1.7	0.9

Note: The protein in bold represent the target protein All0195 from *Anabaena*.

falling into each category. Table 3 shows the number of amino acids present in each category for the proteins present in secondary seed. From the prediction of secondary structure, it is clear that most of the amino acids in the proteins of secondary seed fall in helix region.

Homology modeling and structure validation

The proteins present in secondary seed were individually taken, and protein sequence similarity search was performed using BLASTP against a PDB database. Table 4 shows the identified templates and their percentage

similarity between the individual proteins obtained from secondary seed and their identified template proteins. BLASTP revealed that most of the hypothetical proteins of secondary seed have 3FZ4 protein as their template from *Streptococcus mutans* Ua159 with percentage similarity ranging from 72-85%. For the hypothetical proteins, Syncc9605_0697 of *Synechococcus* CC9605 and Syncc9902_1661 of *Synechococcus* CC9902 have templates from *Staphylococcus aureus* subsp. aureus COL and *Bacteroides fragilis* with 74.3 and 77.4 percentage similarity respectively (table 4). Using these individual protein sequences from the secondary seed along with their identified templates, homology modeling was performed using Modeller version 9.15. The resulting homology models of the primary seed proteins were validated using RAMPAGE server. The Ramachandran plots generated using RAMPAGE server reveals that 94% to 98% amino acid residues of the all the proteins present in the primary seed are in most allowed region (table 5). The model quality was also estimated by super imposition of the model with its template and root mean square deviation (RMSD) was observed for all the protein models of the secondary seed data (table 4). Figure 2 shows the super imposed structures of the models with their templates. The root mean square deviation values obtained by super imposition of the modeled proteins with their respective templates vary from 0.217 Å to 0.512Å (table 4).

Discussion

Arsenic exists in two forms such as oxidized As and reduced As [27]. In cyanobacteria, it was reported that the accumulation of arsenic in the cell leads to the change in levels of chlorophyll α and also has an effect of disorganization in the membranes present in the chloroplasts. As a counter mechanism for the toxicity of arsenic, arsenic resistance pathways were evolved and the genes encoding for the proteins involved in the detoxification arsenic were organized in the form of Operons [28]. Till today the proteins which are reported to have the arsenate reductase properties were classified into three independently evolved families identified in *Escherichia coli*, *Staphylococcus aureus*, and *Saccharomyces cerevisiae* respectively [29]. In addition to these three families, another new hybrid type arsenate reductase was identified in the cyanobacterium *Synechocystis sp.* 6803[26]. Since arsenate reductase is encoded by more than one gene, a fundamental question arises about the other genes which code for arsenate reductase like proteins. In cyanobacteria, it was reported that apart from the genes encoding for *E. coli* like arsenate reductases, the other arsenate reductases have ArsC domain in their primary sequence and belong to thioredoxin super family [30]. The protein encoded by the gene *all0195* of *Anabaena sp.* (reference protein) also has ArsC domain conserved and was also reported that it belongs to thioredoxin family [8]. Moreover, the function of this gene *all0195* was confirmed to be involved in the arsenate detoxification in *Anabaena* by performing two experiments such as complementation assay of arsenate reductase activity in ΔarsC *E. coli* WC3110 and *in vitro* assay of the arsenate reductase activity of purified recombinant All0195 [8]. Transformation of ΔarsC *E. coli* with pGEX-5X-2-all0195 aided the growth of ΔarsC *E. coli* which is nearly similar to that of wild type of *Anabaena* [8]. It was also reported that the *in vitro* assay also gave positive results and was in agreement with the hypothesis that the gene *all0195* codes for arsenate reductase [8].

From our predictions, it is clear that there exists a high similarity in the protein sequences of All0195 and the predicted bidirectional best hits. From the alignment of the protein sequences, it is clear that many of the amino acids, especially the amino acids falling in the region of ArsC domain were well conserved. From conserved

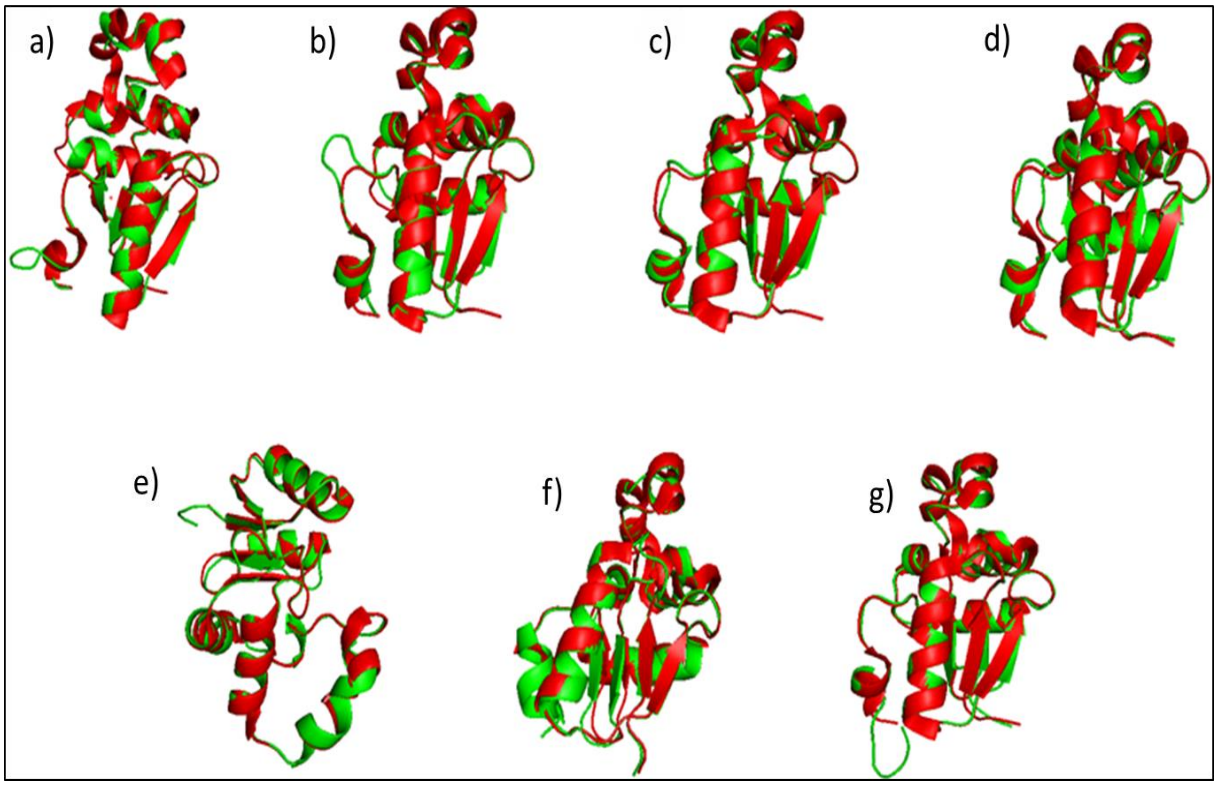

Figure 2. Super imposed structures of modeled Hypothetical proteins (green color) and their templates (red color). The figures a, b, c, d, and g represent the modeled structures (green color) of All0195, Am1_5153, Ava_2687, Pmt9312_0513, Synpcc7002_a0009 with their template 3FZ4 (red color). The figures e and f represent the modeled structures (green color) of Syncc9605_0697 and Syncc9902_1661 with their templates 2M46 and 3GKX respectively.

domain test, it is also clear about the existence of ArsC domain and also its positional conservation in the sequences. By observing the statistical and mathematical values obtained from secondary structures, 3D structures, and Ramachandran plots, it is very promising that the predicted hypothetical proteins which are bidirectional best hits to All0195 may have arsenate reductase properties. It was also very interesting to observe that majority of the bidirectional best hits identified belongs to the cyanobacterial species which live in marine water (table 2). There are also very few bidirectional best hits identified in the cyanobacterial species which have sediment and soil adaptation (table 2). Upon on extensive literature search, we found that the cyanobacteria which are adapted to marine environment and which live in soil and sediment niches are capable of uptake of arsenic [8, 30].

From the reports cited above and the results obtained from the in-depth bioinformatic analysis on the predicted bidirectional best hits of the protein All0195 of *Anabaena sp*., it can be concluded that in this study a total of seven new hypothetical proteins which have putative arsenate reductase protein like properties were computationally identified and in silico characterized.

Reference

1. Whitton BA. 2012. Ecology of cyanobacteria II: their diversity in space and time. Springer Science & Business Media, 65-126.
2. Delwiche CF, Palmer JD. 1997. The origin of plastids and their spread via secondary symbiosis. Springer Vienna, 53-86.
3. Liu H, Nolla HA, Campbell L. 1997. Prochlorococcus growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean. Aquatic Microbial Ecology, 12:39-47.
4. Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY. 1979. Generic assignments, strain histories and properties of pure cultures of cyanobacteria. Microbiology, 111:1-61.

5. Arun PV, Bakku RK, Subhashini M, Singh P, Prabhu NP, Suzuki I, Prakash JS. 2012. CyanoPhyChe: a database for physico-chemical properties, structure and biochemical pathway information of cyanobacterial proteins. PLoS One, 7:e49425.
6. Rosen BP. 2002. Biochemistry of arsenic detoxification. FEBS Lett, 529:86-92.
7. Chaturvedi N, Singh VK, Pandey PN. 2013. Computational identification and analysis of arsenate reductase protein in Cronobacter sakazakii ATCC BAA-894 suggests potential microorganism for reducing arsenate. J Struct Funct Genomics, 14:37-45.
8. Pandey S, Shrivastava AK, Singh VK, Rai R, Singh PK, Rai S, Rai LC. 2013. A new arsenate reductase involved in arsenic detoxification in Anabaena sp. PCC7120. Funct Integr Genomics, 13:43-55.
9. Williams PN, Price AH, Raab A, Hossain SA, Feldmann J, Meharg AA. 2005. Variation in arsenic speciation and concentration in paddy rice related to dietary exposure. Environ Sci Technol, 39:5531-5540.
10. Muller D, Lievremont D, Simeonova DD, Hubert JC, Lett MC. 2003. Arsenite oxidase aox genes from a metal-resistant beta-proteobacterium. J Bacteriol, 185:135-141.
11. Achour AR, Bauda P, Billard P. 2007. Diversity of arsenite transporter genes from arsenic-resistant soil bacteria. Res Microbiol, 158:128-137.
12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol, 215:403-410.
13. Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet, 16:276-277.
14. Ikai A. 1980. Thermostability and aliphatic index of globular proteins. J Biochem, 88:1895-1898.
15. Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. J Mol Biol, 157:105-132.
16. Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol, 302:205-217.
17. Finn RD. et al. 2006. Pfam: clans, web tools and services. Nucleic Acids Res, 34:D247-251.
18. Schultz J, Milpetz F, Bork P, Ponting CP. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A, 95:5857-5864.
19. Chou PY, Fasman GD. 1974. Prediction of protein conformation. Biochemistry, 13:222-245.
20. Chou PY, Fasman GD. 1974. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry, 13:211-222.
21. Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol: 234:779-815.
22. Lovell SC, Davis IW, Arendall WB, 3rd, de Bakker, PI, Word JM, Prisant M.G. , Richardson JS, Richardson DC. 2003. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins, 50:437-450.
23. Arun PPS, Prakash JS. 2016. UpCoT: an integrated pipeline tool for clustering upstream DNA sequences of orthologous genes in prokaryotic genomes. 3 Biotech, 6:1-7.
24. Wolf YI, Koonin EV. 2012. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol Evol, 4:1286-1294.
25. Zhang S, Li S, Niu M, Pham PT, Su Z. 2011. MotifClick: prediction of cis-regulatory binding sites via merging cliques. BMC Bioinformatics, 12:238.
26. Li R, Haile JD, Kennelly PJ. 2003. An arsenate reductase from Synechocystis sp. strain PCC 6803 exhibits a novel combination of catalytic characteristics. J Bacteriol, 185:6780-6789.
27. Oremland RS, Stolz JF. 2005. Arsenic, microbes and contaminated aquifers. Trends Microbiol, 13:45-49.
28. Pandey S, Shrivastava AK, Rai R, Rai LC. 2013. Molecular characterization of Alr1105 a novel arsenate reductase of the diazotrophic cyanobacterium Anabaena sp. PCC7120 and decoding its role in abiotic stress management in Escherichia coli. Plant Mol Biol, 83:417-432.
29. Mukhopadhyay R, Rosen BP. 2002. Arsenate reductases in prokaryotes and eukaryotes. Environ Health Perspect, 110 Suppl 5:745-748.
30. Huertas MJ, Lopez-Maury L, Giner-Lamia J, Sanchez-Riego AM, Florencio FJ. 2014. Metals in cyanobacteria: analysis of the copper, nickel, cobalt and arsenic homeostasis mechanisms. Life (Basel), 4:865-886.