

**Design and development of computational resources for
in silico protein characterization and prediction of
cis-regulatory elements in cyanobacteria**

A Thesis

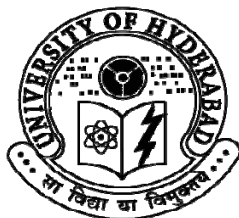
Submitted for the Degree of
DOCTOR OF PHILOSOPHY

ॐ

**V PARVATI SAI ARUN P
09LPPH24**



**DEPARTMENT OF PLANT SCIENCES
SCHOOL OF LIFE SCIENCES
UNIVERSITY OF HYDERABAD
HYDERABAD 500 046
INDIA
April 2015**



School Of Life Sciences,
University Of Hyderabad,
Central University P. O.,
Hyderabad 500 046.
India

DECLARATION

I hereby declare that the matter embodied in this thesis is the result of investigations carried out by me in the Department of Plant Sciences, School of Life Sciences, University of Hyderabad, Hyderabad, under the supervision of **Dr. J. S.S. Prakash**.

In keeping with the general practice of reporting scientific observations, due acknowledgement has been made wherever the work described is based on the findings of other investigators.

Date:

V Parvati Sai Arun P

Place:



School Of Life Sciences,
University Of Hyderabad,
Central University P. O.,
Hyderabad 500 046.
India

CERTIFICATE

Certified that the work embodied in this thesis entitled “**Design and development of computational resources for *in silico* protein characterization and prediction of *cis*-regulatory elements in cyanobacteria**” has been carried out by **V PARVATI SAI ARUN P**, under my supervision. This work is free from plagiarism and has not been submitted elsewhere for a diploma or degree.

Head
(Department of Plant Sciences)

Dr. J. S.S. Prakash
(Thesis Supervisor)

Dean
(School of Life Sciences)

Acknowledgements

I extend my gratitude to my supervisor **Dr.J.S.S.Prakash** for his constant guidance and support throughout my doctoral research. Sir, I am very much thankful to you for all the faith and love you have shown.

I thank **Prof. A.S. Raghavendra**, Dean, School of Life Sciences, and former Dean, **Prof. M. Ramanadham** for allowing me to use the general facilities of the school.

I thank the present and former Heads of the Department of Plant Sciences, **Prof Ch Venkata Ramana**, and **Prof. A.R. Reddy** for allowing me to use the departmental facilities.

I thank my Doctoral Committee members **Prof. P.B.Kirthi** and **Prof. P. Appa Rao** for their valuable suggestions.

I thank **Dr. Sarath Chandra Janga** for his valuable suggestion during the development of *CyanoCis*.

I thank **Prof Norio Murata** for his valuable suggestions in analyzing the *Cis*- regulatory elements data.

I thank **Prof. Iwane Suzuki**, Tsukuba University (Japan) and his suggestions.

My heartfelt thanks to **Dr. Prakash Prabhu** for his constant support during the development of *CyanoPhyChe*.

I thank **DST, CSIR** for providing me financial assistance in the form of fellowship, **DBT** for providing me high end computer.

I thank our lab helpers **Ramesh, Sridhar** for their help.

I thank my seniors **Dr. Sankara Krishna, Dr. Mallikarjun, Dr. Mujahid, Dr. Mahalakshmi,**

Dr. Sirisha, Dr. Sunil, Rajsheel, Y. Srinivas, Dr. Subhashini, Mrs. Bhavani for being with me all the time.

My special thanks to my senior **G. Mahesh kumar, Chandra Mouli.**

I thank my friends **Anil (IOB), Venkat (IIT Madras), Vijay, Neeshat, Teja, Ranjith, Santosh, Uttam, Harish, Deepak, Somashekar, Bantu, Afshan, Jahnabi, Sanjay** for making cheerful atmosphere in the lab.

My special thanks goes to **Anu Madam, Pranav,** and **Chinna** for their kindness and support.

My special thanks goes to **Parikshit, Venkatesh, Father and Mother in laws.**

I thank my parents **P.B Srinivas** and **P.B. Latha** for their eternal love towards me.

I thank my brother **Dr. Kiran** and my sister in law **Dr. Vyshnavi** for being with me all the times.

My heartfelt thanks to my wife **Dr. Syamala** and to my son **Aditya Pavan** for their endless love and making me always happy.

.....Arun.

List of Figures

Figure No	Title	Page No.
1	Home page of Cyanobase	6
2	Snapshot showing the home page of cTFbase	7
3	Snapshot showing the home page of CyanoClust	8
4	Snapshot showing the home page of Cyano2Dbase	9
5	Snapshot showing home page of CyanoEXpress	10
6	Snapshot showing home page of CyanoLyase	11
7	Snapshot showing home page of ProPortal	12
8	Snapshot showing home page of CyanoPhyChe database	23
9	Snapshot of Access page of CyanoPhyChe	24
10	Snapshot of Physico-chemical properties of Sll0649, Sll0088 and Sll0359 proteins.	26
11	SDS-PAGE gel pictures of Sll0649, Sll0088 and Sll0359	27
12	Snapshot showing the CyanoPhyChe Search page	29
13	Snapshot showing home page of ProtPhyChe web server	37
14	Snapshot showing the options provided in the ProtPhyChe	39

	for selection of a organism or to upload file	
15	Snapshot describing the options of Comparative mode of ProtPhyChe	40
16	Snapshot showing the server activity page of ProtPhyChe	41
17	Snapshot showing the partial list of predicted physico-chemical properties of <i>Synechocystis</i> proteome in normal mode	42
18	Snapshot showing the predicted secondary structure of Sds protein of <i>Synechocystis</i>	43
19	Snapshot showing the predicted tertiary structure of Sds protein of <i>Synechocystis</i>	44
20	Snapshot showing the predicted physico-chemical properties of Sds protein of <i>Synechocystis</i> PCC 6803 substr PCC P and its ortholog from <i>Thermosynechococcus elongatus</i> BP-1	44
21	Snapshot showing the super imposed structures of protein 'Solanesyl diphosphate synthase of <i>Synechocystis</i> sp. PCC 6803 substr PCC P and its ortholog from <i>Thermosynechococcus elongatus</i> BP-1.	45

22	Snapshot showing home page of CyanoCis web server	53
23	Snapshot showing the list of genes of cyanobacterium <i>Synechocystis</i> sp. PCC 6803	54
24	Snapshot showing the options for viewing the bidirectional best hits and prediction of <i>cis</i> -regulatory elements	55
25	Snapshot showing list of cyanobacteria and checkboxes for selection of organisms to retrieve bidirectional best hits.	56
26	Snapshot showing the retrieved orthologs for the gene <i>slr2075</i> of <i>Synechocystis</i>	56
27	Snapshot showing the parameter file of CyanoCis.	57
28	Snapshot showing the results page of CyanoCis web server	58
29	Snapshot showing home page of UpCoT	68
30	Snapshot showing the options to select operating system and target organism	69
31	Snapshot showing the list of reference organisms	69
32	Snapshot showing the information at glance in UpCoT	70
33	Snapshot showing the directories and files present in	70

	the UpCoT package	
34	Snapshot showing the file contents of settings.txt file of UpCoT	71
35	The Schematic representation of UpCoT input, UpCoT work flow and UpCoT output	73

List of Tables

Table No	Title	Page No.
1	List of computational resources for aiding cyanobacterial research	4
2	The <i>cis</i> -regulatory elements identified by CyanoCis in the upstreams of selected genes of <i>Synechocystis</i> sp. PCC6803.	60
3	Orthologs identified by UpCoT for selected protein of target organism <i>Synechocystis</i>	76
4	<i>Cis</i> -regulatory elements predicted in the clustered upstreams of selected tgCoTs generated by UpCoT	77

Table of Contents

Chapter 1	General Introduction	1-14
1. Introduction.....		2
1.1 General Information.....		2
1.2 Cyanobacteria and the status of cyanobacterial genome sequencing.....		3
1.3 Databases for aiding cyanobacterial research.....		5
1.3.1 Cyanobase.....		5
1.3.2 cTFbase.....		6
1.3.3 CyanoClust.....		7
1.3.4 Cyano2Dbase.....		8
1.3.5 CyanoEXpress.....		9
1.3.6 CyanoLyase.....		10
1.3.7 MAAs database.....		11
1.3.8 ProPortal.....		11
1.3.9 SynechoNET.....		12
1.4 Web servers for cyanobacterial research.....		13
1.4.1 MUST.....		13
1.5 Objectives.....		14
Chapter 2/Objective 1	CyanoPhyChe	15-30
Summary.....		16
2. Introduction.....		17
2.1. Materials and methods.....		18

2.1.1. Protein sequence data for calculating physico-chemical properties.....	18
2.1.2. Aliphatic index and structural stability.....	19
2.1.3. Grand average value of hydropathy (GRAVY).....	19
2.1.4. Canonical variable for solubility (CVsol) and PEPID.....	20
2.1.5. Secondary structure prediction.....	21
2.1.6 Design of CyanoPhyChe database and its accessibility.....	21
2.1.7 Heterologous expression of <i>Synechocystis</i> proteins in <i>E. coli</i>	22
2.2. Results and Discussion.....	23
2.2.1. Description of CyanoPhyChe.....	23
2.2.2. Browsing the Database.....	23
2.2.3. Physico-chemical properties.....	24
2.2.4. Amino acid composition and structure information.....	27
2.2.5. Biochemical pathway in which a protein is involved.....	28
2.2.6. Search options in the CyanoPhyChe database.....	28
2.3. Conclusion.....	29

Chapter 3/Objective 2 ProtPhyChe 31-46

Summary.....	32
3. Introduction.....	33
3.1. Materials and Methods.....	35
3.1.1. Construction of ProtPhyChe.....	35
3.1.2. User registration.....	35
3.1.3. Selection of target organism.....	35

3.1.4. Normal mode and Comparative mode of <i>in silico</i> characterization.....	35
3.2 Results and discussion	37
3.2.1. Description of ProtPhyChe.....	37
3.2.2. Starting ProtPhyChe.....	38
3.2.3. Selection of a proteins or proteome.....	38
3.2.4. Uploading file containing protein sequences.....	38
3.2.5. Working modes of ProtPhyChe.....	39
3.2.5.1. Normal Mode.....	39
3.2.5.2. Comparative Mode.....	39
3.2.6. Format of results.....	41
3.3. Applications of ProtPhyChe.....	42
3.3.1. Normal mode application.....	42
3.3.2. Comparative mode application.....	43
3.4 Conclusion.....	45
Chapter 4/Objective 3	CyanoCis
	47-62
Summary.....	48
4. Introduction.....	49
4.1. Phylogenetic footprinting.....	49
4.2. Methods	
4.2.1. Selection of cyanobacterial genomes.....	50
4.2.2. Identification of orthologs and generation of cyCoGs.....	51

4.2.3. Clustering upstream DNA sequences of genes of a cyCoT.....	51
4.2.4. Parameters used for finding <i>cis</i> -regulatory elements for selected <i>Synechocystis</i> genes.....	53
4.3. Results and Discussion.....	53
4.3.1. Description of CyanoCis.....	53
4.3.2. Interface of the CyanoCis.....	54
4.3.3. Identification of orthologs of selected cyanobacterial genes.....	55
4.3.4. Prediction of <i>Cis</i> -regulatory elements.....	55
4.4. Output of CyanoCis.....	58
4.5. Performance analysis of CyanoCis.....	59
4.6. Conclusion.....	62
Chapter 5/Objective 3	UpCoT
	63-77
Summary.....	64
5. Introduction.....	66
5.1. Materials and Methods.....	67
5.1.1. Design of UpCoT interface.....	67
5.2. Description and accessibility of UpCoT web interface.....	67
5.3. UpCoT output.....	70
5.4. Methodology used for testing UpCoT.....	72
5.5. Results and Discussion.....	73
5.5.1. Performance analysis of UpCoT.....	73
5.5.2. Analysis of clustered-upstream DNA sequences for selected tgCoTs.....	74

5.6. Conclusion.....	75
----------------------	----

References	76-84
-------------------	--------------

Chapter 1

General Introduction

1 . Introduction

1.1 General Information

In recent years, high throughput genome sequencing became economically feasible, delivering the bacterial sequences in hours to days (Loman, et al., 2012). The advancements in genome sequencing technologies made the sequencing of human genome with a low price of \$5000 to \$10000 (Loman et al., 2012; Stahl and Lundeberg, 2012). With the advantage of fast and new generation automated DNA sequencing technologies, a number of microbial genomes have been sequenced during the past decade and the sequence information is available in various genome databases. The researchers, those are interested in understanding a wide range of topics related to bacterial genetics, genomics, molecular biology, adaptation studies and molecular evolution, opt for sequencing of the bacterial genome in question (Edwards and Holt, 2013). The genome sequencing technology became very handy such that genome sequences of many bacteria are generated in their own labs by many research groups in a matter of few hours to days using bench top sequencers such as the Illumina MiSeq, Ion Torrent PGM or Roche 454 FLX Junior etc (Loman et al., 2012; Stahl and Lundeberg, 2012). Using such fast and automated DNA sequencing technologies, the public databases such as Genbank has been updated with more than 6500 bacterial genome assemblies, of which about two thirds are in draft form by february 2013 (Edwards and Holt, 2013). These large datasets, which are available in the public repositories can be used for different studies such as comparative and functional genomics, phylogenetics, adaptation, and molecular evolution.

1.2 Cyanobacteria and the status of cyanobacterial genome sequencing

Cyanobacteria are ancient photosynthetic bacteria and are widely distributed in different environments ranging from aquatic, hot springs, deserts, and polar environments (Whitton,

2012). They have wide morphological differences and exist as unicellular to filamentous forms (Whitton, 2012). They are considered as the globally important primary producers (Rosmarie Rippka, 1978; Hongbin Liu, 1997) and are also considered to be the progenitors of higher plant chloroplasts (Delwiche and Palmer, 1997). Some diazotrophic cyanobacteria are reported to be responsible for global nitrogen fixation and therefore play a significant role in the nitrogen fixation, as well as carbon and oxygen balancing (John, 1998). As cyanobacteria are adapted to different environments and possess vital metabolic pathways, they became good model systems for studies related to adaptation, abiotic stress, production of secondary metabolites, gene expression etc. Genome sequencing of cyanobacteria was first initiated in the year 1996. The first cyanobacterial genome sequenced is of cyanobacterium *Synechocystis* sp. PCC 6803 (Kaneko et al., 1996). From 1996 to till today there are about 98 cyanobacterial genomes were completely sequenced and are publicly available at Genbank (<http://www.ncbi.nlm.nih.gov/genbank/>). Using this genome data of cyanobacteria and by using functional genomics and proteomics approaches several questions were answered in previous reports related to evolution, adaptation, physiology, and biochemistry of cyanobacteria. Using bioinformatics approaches researchers have developed few computational resources and tools for aiding research on cyanobacteria (**Table 1**). These databases and web servers were developed for a single cyanobacterial species such as CyanoExpress (Hernandez-Prieto and Futschik, 2012) as well as for group of cyanobacteria, such as CyanoClust (Sasaki and Sato, 2010). Description of databases, which are available for the aid of research on cyanobacteria are given in **Table 1**.

Table 1: List of computational resources for aiding cyanobacterial research

Name of the resource	Information	Reference
Cyanobase http://genome.microbedb.jp/cyanobase/	Cyanobacterial genome database.	(Nakao et al., 2010)
cTFbase http://bioinformatics.zj.cn/cTFbase/index.php	Provides the information of transcription factor repositories in cyanobacteria.	(Wu et al., 2007)
Cyanoclust http://cyanoclust.c.u-tokyo.ac.jp/	Comparative genomics database which provides information about the orthologs, among the cyanobacteria and plastids.	(Sasaki and Sato, 2010)
Cyano2Dbase http://wiki.annotation.jp/Cyano2Dbase .	Provides the information about the protein-gene linkage maps of <i>Synechocystis</i> sp. PCC 6803.	(Sazuka et al., 1999)
CyanoEXpress http://cyanoexpress.sysbiolab.eu/	Database developed for exploration and visualization of transcription profiles in <i>Synechocystis</i> sp. PCC 6803.	(Hernandez-Prieto and Futschik, 2012)
CyanoLyase http://cyanolyase.genouest.org/	Database of phycobilin lyases and related protein sequences.	(Bretaudeau et al., 2013)
MAAs http://www.biologie.uni-erlangen.de/botanik1/html/eng/maa_data_base.htm	Provides the information about UV-absorbing microsporines and microsporine like amino acids present in fungi, cyanobacteria, phytoplankton, macro-algae, and animals.	(Sinha et al., 2007)
ProPortal http://proportal.mit.edu/	Serves as a source genomic, metagenomic, transcriptomic and field data of <i>Prochlorococcus</i> , <i>Synechococcus</i> .	(Kelly et al., 2012)

SynechoNet http://biportal.kobic.re.kr/SynechoNET/	Provides information about protein-protein interaction useful for analysis of regulatory membrane proteins in cyanobacteria <i>Synechocystis</i> .	(Kim et al., 2008)
MUST http://csbl1.bmb.uga.edu/ffzhou/	System providing the information about the novel miniature inverted transposable elements.	(Chen et al., 2009)

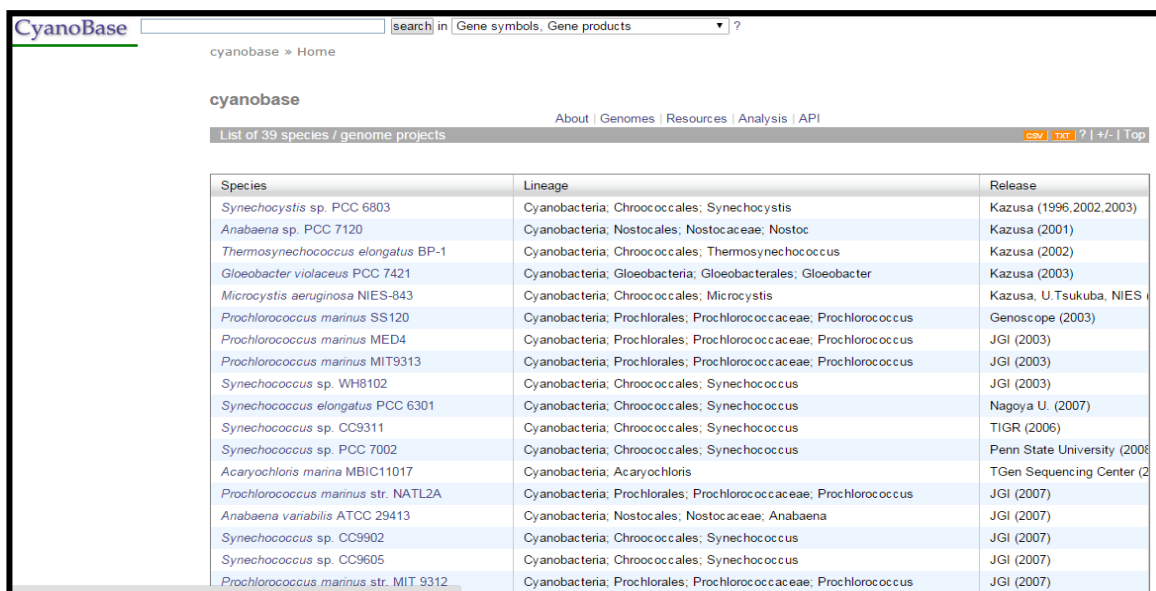
1.3 Databases for aiding cyanobacterial research

There are a several databases available for aiding research on cyanobacteria. There are about nine databases and one web server exists till date aiding cyanobacterial research (**Table 1**).

1.3.1 Cyanobase

Cyanobase is a genome database of cyanobacteria. The first version of Cyanobase was developed in the year 1998 (Nakamura et al., 1998). It contained the cyanobacterial genome sequence information, gene annotation, functional classification of cyanobacterial genes, location of the genes on the genome, nucleotide sequences of the genes and deduced amino acid sequences. Later in the year 1999, as an extension to Cyanobase, a repository database named as CyanoMutants was included into Cyanobase (Nakamura et al., 1999). This CyanoMutants repository includes the information of *Synechocystis* mutants. In this version of Cyanobase, each entry in the Cyanobase was modified in such a way that it contained data describing about the gene identifier and information about availability of the mutant. The third version of the Cyanobase was released in the year 2000 (Nakamura et al., 2000). In this release, database was updated with the experimental information for the genes, which were putative in nature. The latest update of Cyanobase was done in the year 2010 (Nakao et al., 2010). **Figure 1** show the home page of the latest version of Cyanobase. This version of

Cyanobase includes the information about 35 completely sequenced cyanobacterial genomes. Further, it also includes the information about various genome scale experiments, gene expression profiles and protein-protein interaction data. This version is completely redesigned with improved accessibility and better organization of the data (**Figure 1**).



Species	Lineage	Release
<i>Synechocystis</i> sp. PCC 6803	Cyanobacteria; Chroococcales; Synechocystis	Kazusa (1996,2002,2003)
<i>Anabaena</i> sp. PCC 7120	Cyanobacteria; Nostocales; Nostocaceae; Nostoc	Kazusa (2001)
<i>Thermosynechococcus elongatus</i> BP-1	Cyanobacteria; Chroococcales; Thermosynechococcus	Kazusa (2002)
<i>Gloeobacter violaceus</i> PCC 7421	Cyanobacteria; Gloeobacteria; Gloeobacterales; Gloeobacter	Kazusa (2003)
<i>Microcystis aeruginosa</i> NIES-843	Cyanobacteria; Chroococcales; Microcystis	Kazusa, U. Tsukuba, NIES
<i>Prochlorococcus marinus</i> SS120	Cyanobacteria; Prochlorales; Prochlorococcaceae; Prochlorococcus	Genoscope (2003)
<i>Prochlorococcus marinus</i> MED4	Cyanobacteria; Prochlorales; Prochlorococcaceae; Prochlorococcus	JGI (2003)
<i>Prochlorococcus marinus</i> MIT9313	Cyanobacteria; Prochlorales; Prochlorococcaceae; Prochlorococcus	JGI (2003)
<i>Synechococcus</i> sp. WH8102	Cyanobacteria; Chroococcales; Synechococcus	JGI (2003)
<i>Synechococcus elongatus</i> PCC 6301	Cyanobacteria; Chroococcales; Synechococcus	Nagoya U. (2007)
<i>Synechococcus</i> sp. CC9311	Cyanobacteria; Chroococcales; Synechococcus	TIGR (2006)
<i>Synechococcus</i> sp. PCC 7002	Cyanobacteria; Chroococcales; Synechococcus	Penn State University (2006)
<i>Acaryochloris marina</i> MBIC11017	Cyanobacteria; Acaryochloris	TGen Sequencing Center (2006)
<i>Prochlorococcus marinus</i> str. NATL2A	Cyanobacteria; Prochlorales; Prochlorococcaceae; Prochlorococcus	JGI (2007)
<i>Anabaena variabilis</i> ATCC 29413	Cyanobacteria; Nostocales; Nostocaceae; Anabaena	JGI (2007)
<i>Synechococcus</i> sp. CC9902	Cyanobacteria; Chroococcales; Synechococcus	JGI (2007)
<i>Synechococcus</i> sp. CC9605	Cyanobacteria; Chroococcales; Synechococcus	JGI (2007)
<i>Prochlorococcus marinus</i> str. MIT 9312	Cyanobacteria; Prochlorales; Prochlorococcaceae; Prochlorococcus	JGI (2007)

Figure 1: Home page of Cyanobase released in the year 2010.

Features like viewing genome resources, genome projects (Dataset view), individual genome project (Species View), Individual chromosomes (Map View), Genes View, Gene category view and Word clouds were incorporated in this version. Cyanobase can be accessed by browsing the link <http://genome.microbedb.jp/cyanobase/>

1.3.2 cTFbase

To better understand the regulatory mechanisms and to build the gene regulatory network of an organism, identification of sensory proteins, signal transducers, transcriptional regulators, regulons of each transcription factor and their target binding sites are important. To provide insights into transcription factor repertoires of the cyanobacteria cTFbase was developed in 2007 (Wu et al., 2007). cTFbase includes total 1288 putative transcription factors, which were

identified from 21 cyanobacterial genomes. We can download the protein sequences, domain architecture information, and annotation of cyanobacterial transcription factors from cTFbase. cTFbase also provides the information on relatedness among various transcription factors in the form of phylogenetic trees. It generates multiple sequence alignments and identifies orthologs for a selected transcription factor among selected cyanobacteria. cTFbase can be accessed through the link <http://bioinformatics.zj.cn/cTFbase/index.php>

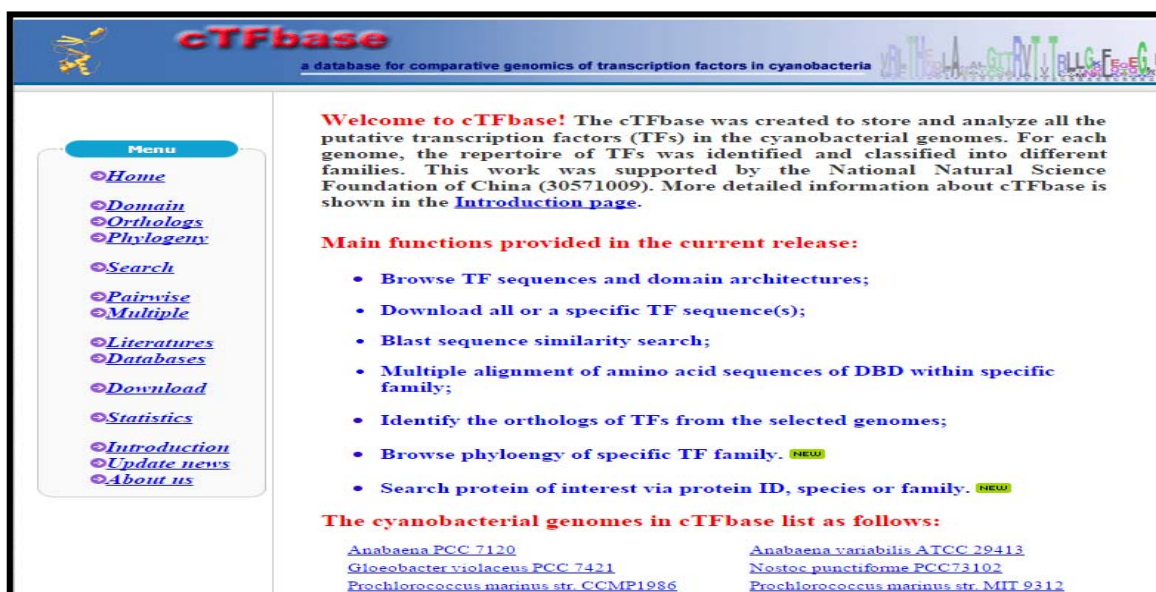


Figure 2: Snapshot showing the home page of cTFbase database.

1.3.3 CyanoClust

CyanoClust is a database developed in the year 2009 (Sasaki and Sato, 2010). CyanoClust includes protein homology information of about 38 cyanobacteria, 59 plastids and 5 anoxygenic photosynthetic bacteria. It also contains the homology information of 'chromatophore' of an ameboid *Paulinella*. CyanoClust includes a total of 1,79,056 proteins which were clustered into 40,526 clusters. CyanoClust can be accessed from <http://cyanoclust.c.u-tokyo.ac.jp/>

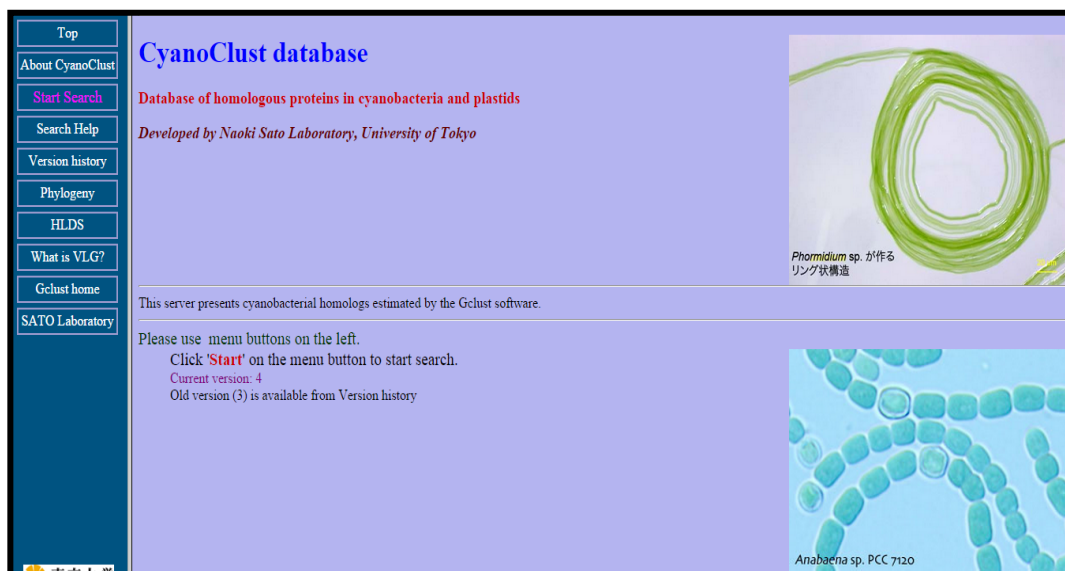


Figure 3: Snapshot showing the home page of CyanoClust database.

1.3.4 Cyano2Dbase

A proteome project on *Synechocystis* was initiated for understanding the protein-gene linkages in *Synechocystis* (Sazuka and Ohara, 1997). In this project, a protein-gene linkage map of *Synechocystis* was constructed for 130 high abundance proteins present on 2D gels. Using these 130 protein spots, the authors attempted to link the protein spots with the genes encoding them (Sazuka and Ohara, 1997). As an extension of this work, the authors attempted to analyze the 2D-gel protein spots from soluble, insoluble, thylakoid membrane, and secretory protein fractions of *Synechocystis* and linked these identified proteins with the genes encoding them. Subsequently, using this data, Cyano2Dbase, a web resource was constructed that provides protein-gene linkage maps of the cyanobacterium *Synechocystis* sp. PCC 6803 (Sazuka et al., 1999). The resultant protein-gene linkage maps were increased to a total of 227 protein spots. The user of the Cyano2Dbase can extract information related to translation, post translational modifications proteins and signal sequences in proteins of *Synechocystis*. Cyano2Dbase can be accessed from <http://wiki.annotation.jp/Cyano2Dbase>.

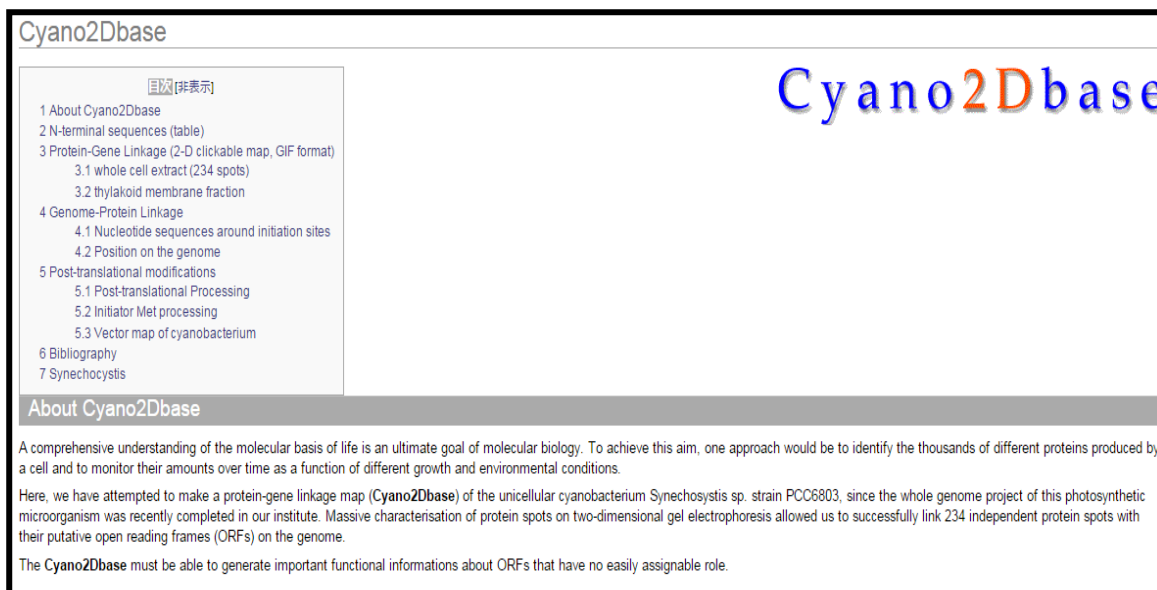


Figure 4: Snapshot showing the home page of Cyano2Dbase database.

1.3.5 CyanoEXpress

CyanoEXpress is an interactive database developed for exploration and visualization of transcriptional response patterns of *Synechocystis* (Hernandez-Prieto and Futschik, 2012). CyanoEXpress database is currently composed with the gene expression data of about 3073 genes and 178 environmental and genetic manipulations obtained from 31 independent studies. The home page of CyanoEXpress is shown in **Figure 5**. A total of 645 genome-wide microarray experimental data carried in 31 independent studies were evaluated and incorporated in CyanoEXpress. The CyanoEXpress provides the information on gene expression profiles for various experimental conditions or genetic manipulations. This can also be used for identification of coherent expression patterns of the genes in a particular condition which implies the identification of co-regulated genes. Moreover it can be used for annotating the genes with unknown function basing on the co-expression pattern in a given condition. CyanoExpress can be accessed from the link <http://cyanoexpress.sysbiolab.eu/>



Figure 5: Snapshot showing the home page of CyanoExpress database.

1.3.6 CyanoLyase

CyanoLyase is a database containing the manually curated sequence and motifs of phycobilin lyases and related proteins (Bretaudiere et al., 2013). Phycobilin enzymes are very important enzymes involved in the covalent ligation of Phycobilins to specific binding sites of phycobiliproteins. These phycobiliproteins are the building blocks of the major light harvesting components, phycobilisomes. Phycobilin lyases are important precursors in the formation of phycobilisomes. As phycobilin lyase sequences are poorly annotated in the public databases, the developers of CyanoLyase classified phycobilin lyases into 3 clans and 32 families and constructed CyanoLyase database. The home page of CyanoLyase is as shown in **Figure 6**. CyanoLyase was developed in such a way that it can annotate the lyases from any newly sequenced genome upon its submission. It provides the phylogenetic profiles of all the phycobilin lyases families, describes their function and also gives the information of their conservation across the genomes present in the database. The overall goal behind developing CyanoLyase was to provide the researcher an extensive collection of information about

phycobilin lyases their classification into clans, subclans, families, subfamilies, to make ease of annotation of the sequences from forthcoming genomes of phycobiliproteins. CyanoLyase can be accessed from <http://cyanolyase.genouest.org/>.



Figure 6: Snapshot showing the home page of the CyanoLyase database.

1.3.7 MAAs database

MAAs database is the database exclusively developed on microsporines and microsporine like amino acids (Sinha et al., 2007). This database includes the information about the UV-absorbing microsporines and microsporine like amino acids present in fungi, cyanobacteria, phytoplankton, macroalgae, and animals from aquatic and terrestrial habitats. This database includes the information of absorption maxima and molecular structures of various mycosporines and mycosporine like amino acids. MAA's database can be accessed from http://www.biologie.uni-erlangen.de/botanik1/html/eng/maa_database.htm

1.3.8 ProPortal

ProPortal database was developed to provide the cross reference data covering from genome to ecosystem of the marine cyanobacterium *Prochlorococcus*, and *Synechococcus*.

It also includes the information of the phage that infects them (Kelly et al., 2012). ProPortal serves as a source of genomic, metagenomic, transcriptomic and field data of *Prochlorococcus*.

Figure 7 show the home page of ProPortal. ProPortal can be accessed from <http://proportal.mit.edu/>

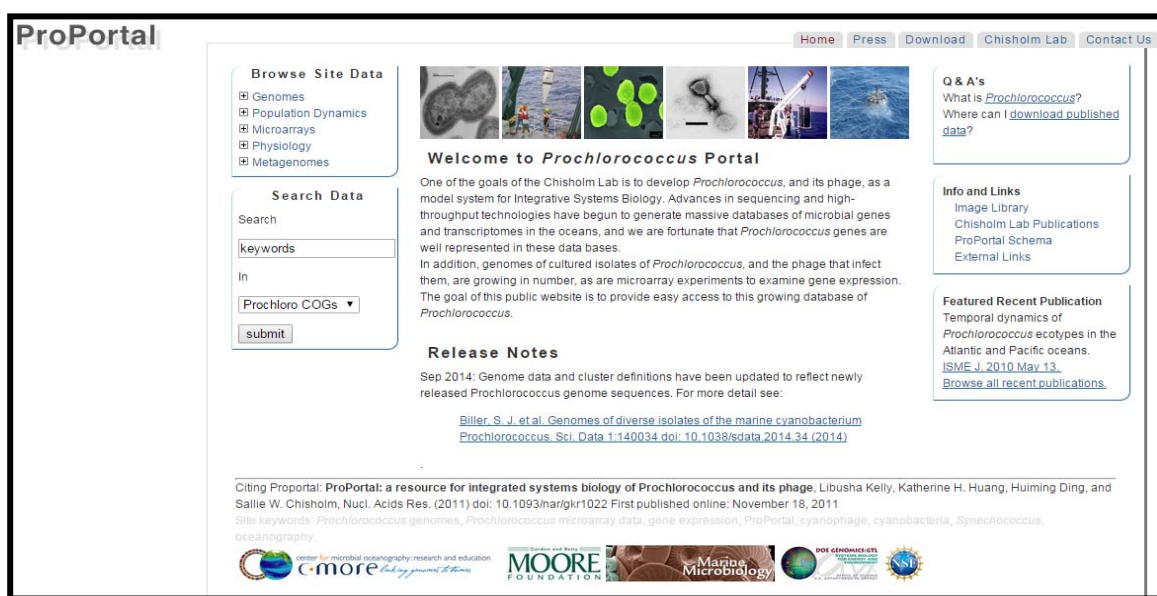


Figure 7: Snapshot showing the home page of ProPortal web resource.

1.3.9 SynechoNET

SynechoNET contains protein-protein interaction data, which is useful for analysis of regulatory membrane proteins in *Synechocystis* (Kim et al., 2008). SynechoNET was developed by integrating four public protein-protein interaction databases such as 'Protein structural interactome map (PSIMAP), iPfam, InterDom, and STRING (Park et al., 2001; Ng et al., 2003; von Mering et al., 2007; Finn et al., 2014). It is designed for predicting trans-membrane topology and domain structure of the regulatory membrane proteins thereby providing the platform for the researchers to extend the genomic data of cyanobacteria for understanding the interaction partners, membrane association and membrane topology of *Synechocystis* proteins. The web accessible link of SynechoNET is <http://biportal.kobic.re.kr/SynechoNET/>

1.4 Web servers for cyanobacterial research

A web server named as "MUST", was developed for identification of miniature inverted repeat transposable elements in the genomes of a cyanobacterium *Anabaena variabilis* and an archaean, *Haloquadratum walsbyi* (Chen et al., 2009). The brief description of MUST web server is given below.

1.4.1 MUST

Transposable elements play crucial part in evolution of genes and genomes (Bureau and Wessler, 1994; Palomeque et al., 2006; Mason Gamer, 2007). Miniature inverted transposable elements are a class of small transposable elements present high in number in an host genome (Casacuberta and Santiago, 2003). Identification of all the miniature inverted transposable elements in a genome could provide new insights about gene evolution and genome dynamics of an organism. MUST (Miniature Inverted transposable elements uncovering system) was developed for prediction and analyses of Miniature Inverted transposable elements at a whole genome level (Chen et al., 2009). MUST could identify known and novel miniature inverted transposable elements when tested on the genome of *Anabaena variabilis* ATCC 29413 and *Haloquadratum walsbyi* DSM 16790. MUST is available at <http://csbl1.bmb.uga.edu/ffzhou/>.

In addition to these existing computational resources, we developed four computational resources for aiding cyanobacterial research. Out of these four resources two are web servers, one database and a software package. We named these computational resources as CyanoPhyChe, ProtPhyChe, CyanoCis and UpCoT. CyanoPhyChe is a database that provides the information about physico-chemical properties, structure and biochemical pathway information of cyanobacterial proteins. ProtPhyChe is a web server, which is an extension for

CyanoPhyChe, capable of performing *in silico* characterization of prokaryotic proteins. CyanoCis is a web tool for prediction of *cis*-regulatory elements in cyanobacterial genomes. UpCoT is a software for automation of first step of Phylogenetic footprinting. We took the design and development of these four computational resources as four objectives. These computational resources developed are completely automated and user friendly in nature. Each of these computational resources are described in detail as individual chapters.

1.5 Objectives

The objectives of the thesis were to design and develop the following computational resources.

1. CyanoPhyche, A database for physico-chemical properties, structure and biochemical pathway information of cyanobacterial proteins.
2. ProtPhyChe, A web server for *in silico* characterization of prokaryotic proteins.
3. CyanoCis, A web tool for the identification of *cis*-regulatory elements in cyanobacterial genomes.
4. UpCoT, An integrative pipeline tool for clustering upstream DNA sequences of orthologous genes in prokaryotic genomes.

Chapter 2

Objective 1

CyanoPhyChe: A Database for Physico-Chemical Properties, Structure and Biochemical Pathway Information of Cyanobacterial Proteins.

CyanoPhyChe: A Database for Physico-Chemical Properties, Structure and Biochemical Pathway Information of Cyanobacterial Proteins.

Summary:

In this chapter, the database CyanoPhyChe has been described. It is a user friendly database that one can browse through for physico-chemical properties, structure and biochemical pathway information of cyanobacterial proteins. We downloaded all the protein sequences from the cyanobacterial genome database for calculating the physico-chemical properties, such as molecular weight, net charge of protein, isoelectric point, molar extinction coefficient, canonical variable for solubility, grand average hydropathy, aliphatic index, and number of charged residues. Based on the physico-chemical properties, we provide the polarity, structural stability and probability of expressed protein entering into an inclusion body (PEPIB). We used the data generated on physico-chemical properties, structure and biochemical pathway information of all cyanobacterial proteins to construct CyanoPhyChe. The data can be used for optimizing methods of expression and characterization of cyanobacterial proteins. Moreover, the ‘Search’ and data export options provided will be useful for proteome analysis. Secondary structure was predicted for all the cyanobacterial proteins using PSIPRED tool and the data generated is made accessible to researchers working on cyanobacteria. In addition, external links are provided to biological databases such as PDB and KEGG for molecular structure and biochemical pathway information, respectively. External links are also provided to different cyanobacterial databases. CyanoPhyChe can be accessed from the following URL: <http://bif.uohyd.ac.in/cpc>.

2. Introduction

As proteins mediate and control coordinated biochemical transformations and cellular processes that are central to activity of life forms, characterization of proteins provide insights into the structure and function of a cell (Creighton, 1993). Pure form of a protein is needed for its characterization and can be extracted through expression and purification techniques. Instead of getting an active and soluble form of a protein, there are chances for a protein to enter into an inclusion body (Fink, 1998; Kopito, 2000). Obtaining functionally active protein from an inclusion body requires denaturation of the protein, followed by refolding into its native form. This is a slow and difficult process which greatly reduces the net yield and activity of the protein (Harrison, 2000). Therefore, prevention of a protein to enter into inclusion body is better than resolving it. Solubility of a protein depends on its physico-chemical properties. For instance, length of a protein, composition and properties of amino acid residues in a protein influence its solubility (Wilkinson and Harrison, 1991). Folding of an expressed protein also depends on the conditions employed during the process of expression and purification. It is possible to prevent the aggregation of a protein by providing suitable conditions based on its physico-chemical properties. The physico-chemical properties can be used to predict the nature of a protein and this information is useful for optimization of expression methods. These properties help to understand the native environment in which the protein will be in soluble and active form, thus aids the researchers in the characterization studies on the proteins of interest. In addition to the available traditional standard laboratory methods for determining physico-chemical properties of a protein, mathematical methods have also been in use for calculating the same based on primary sequence information (Levene PA, 1923; Chou and Fasman, 1978; Ikai, 1980; Kyte and Doolittle, 1982; Gill and von Hippel, 1989; Wilkinson and Harrison, 1991; Nakashima and Nishikawa, 1994; Idicula-Thomas and Balaji, 2005). Though different web based tools are

available to determine physico-chemical properties of proteins (McGuffin et al., 2000; Rice et al., 2000; Mathura and Kolippakkam, 2005; Li et al., 2006; Wishart et al., 2008), it is time consuming and difficult task for a naive user in choosing a tool among the available pool. Moreover there is no database, which can provide physico-chemical properties of cyanobacterial proteins. This motivated us to make a database on physico-chemical properties of cyanobacterial proteins using different tools and formulae which are scientifically proven to be accurate (Ikai, 1980; Kyte and Doolittle, 1982; Wilkinson and Harrison, 1991; McGuffin et al., 2000; Rice et al., 2000; Wishart et al., 2008). Thus, we designed a user friendly database in which one can easily search for the properties of a cyanobacterial protein(s) in question and get preliminary understanding about them. We used the genome data from the database of cyanobacteria for calculating the physico-chemical properties of all cyanobacterial proteins and generated the database 'CyanoPhyChe'. Researchers can use the physico-chemical properties of any cyanobacterial protein that is available in the database for their research. The information provided in the database aids researchers for choosing optimal conditions for expression, purification and characterization of a cyanobacterial protein. Moreover, user can export the physico-chemical properties, predicted secondary structure, amino acid sequence and amino acid composition of selected cyanobacterial proteins for further analysis. External links are provided to make a direct access to other cyanobacterial databases available on the internet.

2.1 Materials and Methods

2.1.1 Protein sequence data for calculating physico-chemical properties

Total 38 files with extension.faa, each containing all protein sequences of a cyanobacterial species, were downloaded from NCBI (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria>).

A code was developed in Perl to separate all the protein sequences of a '.faa' file into multiple FASTA files, each with an individual protein sequence. This primary seed data is used for calculating protein properties. Seed data contained total 1,26,610 proteins, covering 38 cyanobacteria. Certain physico-chemical properties were calculated using PEPSTATS tool, which is available in EMBOSS package, installed in a computer using Linux mint-12 operating system(http://www.ebi.ac.uk/Tools/seqstats/emboss_pepstats/). PEPSTATS provides molecular weight, number of residues, isoelectric point (pI), molar extinction coefficient and amino acid composition of a protein (Rice et al., 2000). The output file generated by PEPSTATS was used as a secondary seed data for calculating other properties like, aliphatic index (AI), GRAVY, canonical variable for solubility (CVsol) and probability of expressed protein entering into an inclusion body (PEPIB).

2.1.2 Aliphatic index and structural stability

Stability of a protein can be calculated using aliphatic index. We used the formula that was developed for determining aliphatic index (Patrick Argos 1979; Ikai, 1980). The aliphatic index values of cyanobacterial proteins were normalized between 0 and 10 to predict the structural stability.

2.1.3 Grand average value of hydropathy (GRAVY)

An empirical formula for calculating hydropathy value for a protein was developed by Kyte and Doolittle in 1982, wherein the hydrophilic and hydrophobic properties of each amino acid side chain in a protein are taken into consideration (Kyte and Doolittle, 1982). Positive hydropathy value is indicative of a polar protein and negative value indicates a non-polar protein.

2.1.4 Canonical variable for solubility (CVsol) and PEPiB

A mathematical formula, based on the amino acid composition and their properties, was derived to predict the solubility of a protein and its probability to enter into an inclusion body (Wilkinson and Harrison, 1991). The solubility or insolubility of a protein can be predicted from the canonical variable, which is a composite parameter of cysteine fraction, proline fraction, turn forming residue fraction, approximate charge average, number of residues and hydrophilicity, according to Wilkinson and Harrison model (Wilkinson and Harrison, 1991). The formula for calculating the canonical variable is given below.

$$\text{Canonical variable (CV)} = \lambda_1 \left(\frac{N + G + P + S}{n} \right) + \lambda_2 \left| \frac{(R + K) - (D + E)}{n} - 0.03 \right|$$

where,

n = number of amino acids in protein

N, G, P, S = number of Asn, Gly, Pro & Ser residues respectively.

R, K, D, E = number of Arg, Lys, Asp & Glu residues.

λ_1 and λ_2 = Coefficients (15.43 and -29.56 respectively) and

Canonical variable for solubility (CVsol) = (CV - CV')

Canonical variable for solubility where, CV' = 1.71.

Probability of solubility or insolubility = $0.4934 + 0.276|CV - CV'| - 0.0392(CV - CV')^2$.

The formula for calculating probability of solubility or insolubility for any protein when expressed in *E. coli* was further evaluated by Harrison in the year 2000 (Harrison, 2000). Also, a detailed description of canonical variable and its usage has been given by Koschorreck et al. 2005 (Koschorreck et al., 2005). Based on the sign of CVsol value PEPiB was determined. If CVsol value is positive, the above equation provides the probability of insolubility, which is considered as PEPiB. If CVsol value is negative, the equation provides the probability of solubility (Harrison, 2000). PEPiB of these proteins were calculated by converting their probability of solubility into the probability of insolubility, considering the fact that the sum of probability of solubility and insolubility must be one.

2.1.5 Secondary Structure Prediction

Secondary structure prediction was done using PSIPRED tool (McGuffin et al., 2000). UNIREF90 database was used as target for PSI-Blast (<ftp://ftp.ebi.ac.uk/pub/databases/uniprot/uniref/>). A BASH shell program was designed to automate this tool for predicting secondary structure for cyanobacterial proteins.

2.1.6 Design of CyanoPhyChe database and its accessibility

The CyanoPhyChe database was developed using MySQL database management system (MySQL Version 5.1.4.1 and php MyAdmin 3.2.4). Web interface was designed using HTML, Java script and PHP to retrieve and visualize the data. PDB IDs and biochemical pathway IDs were extracted from NCBI (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria>) and KEGG (<http://www.genome.jp/kegg/>) databases respectively for all cyanobacterial proteins. The pathway IDs were provided with links using HTML and PHP scripting. CyanoPhyChe can be accessed from a local web server located in Bioinformatics Infrastructure Facility at School of

life Sciences, University of Hyderabad, India using the following URL:
<http://bif.uohyd.ac.in/cpc>.

2.1.7 Heterologous expression of *Synechocystis* proteins in *E. coli*

Open reading frames coding for Sll0649, Sll0088 and Sll0359 were amplified, by PCR, with the primer sets Sll0649-F: 5' -GAC TCA TAT GTG GGG GAA CAG GAC TGA A -3' and Sll0649-R: 5' - GCT GAA TTC TTA ATC AGG GTC TTC AAA CTT AT-3', Sll0088-F: 5'-GAC TCA TAT GGG GGT TGT CCT TTC AGT T -3' and Sll0088-R: 5' -GCT GAA TTC TTA GTT CGG GGT TTT AGA CTG G -3' and Sll0359-F: 5' -GAC TCA TAT GCC AAA CGC CTC CAC CGC-3' and Sll0359-R: 5' -GCT GAA TTC TTA TAC TTC CTC TTC GTC ATC G -3'. The amplified ORFs were eluted from the gel and were inserted into pET-28a(+) at suitable restriction enzyme sites to generate pET-Sll0649, pET-Sll0088 and pET-Sll0359. The N-terminally His-tagged proteins were expressed in BL21(DE3)pLysS, which had been transformed with the above DNA constructs. The expression of each protein was induced by addition of 0.4 mM (final concentration) IPTG. Bacterial cells were collected by centrifugation at 10 000 g for 10 min and pelleted cells were disrupted with a sonic oscillator (model, UV2070; probe, MS-72; Bandelin Electronic) operated for 10 min at 50% power, with 1 min pulse interval, in 100 mM Tris/HCl (pH 8.0) and 200 mM NaCl. Soluble supernatant and insoluble precipitates were separated by centrifugation at 20,000 g for 20 min at 4°C. Insoluble fractions were suspended in the same buffer which contained a 0.5% Triton -X 100. The soluble and insoluble proteins were resolved on 12% SDS-PAGE.

2.2. Results and Discussion

2.2.1 Description of CyanoPhyChe

The web interface contains 'Home', 'Browse', 'Search', 'Help' and 'Contact' links on the top left corner of the index page for browsing the database (**Figure 8**). A brief introduction about CyanoPhyChe is given in the home page. The physico-chemical properties, structural and biochemical pathway information of any cyanobacterial protein can be accessed either by 'Browse' or 'Search' link. Further, the web page contains external links to other cyanobacterial databases such as cyanobase, cyanoclust, cyanoDB and cyanoBIKE. This will assist the user to get a direct access to other related databases for additional information on cyanobacteria.

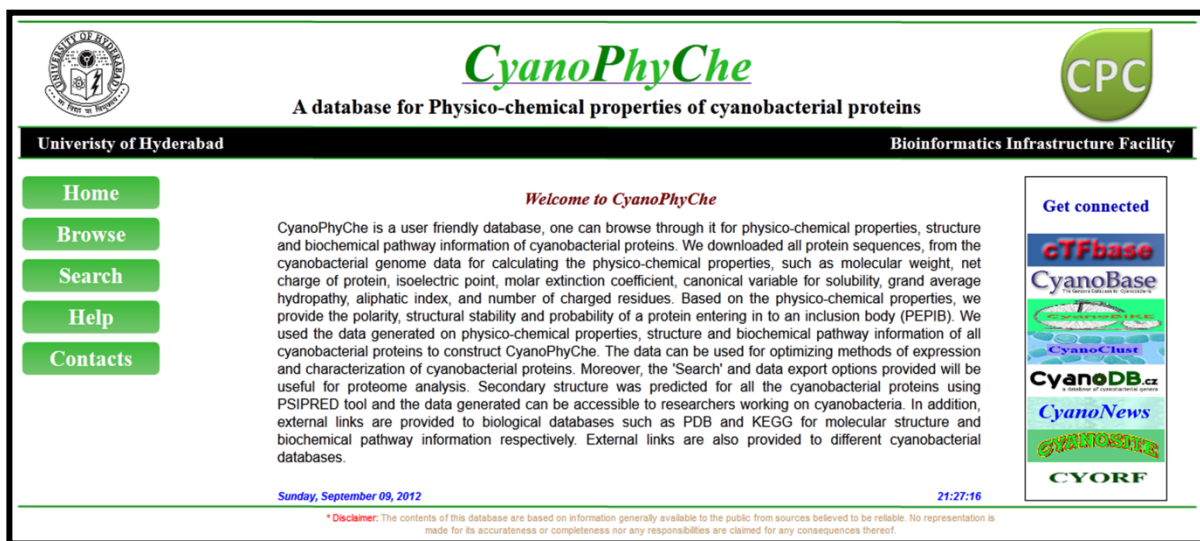


Figure 8: Snapshot of the homepage of CyanoPhyChe database. User can navigate for physico-chemical properties of any cyanobacterial protein by browsing through 'Browse' link located at the top left corner of the home page. Proteins with specific properties can be retrieved using 'Search' option. Links to external cyanobacterial databases are provided in the main page of the database.

2.2.2 Browsing the Database

As the user goes through the 'Browse' option, the list of cyanobacteria is displayed. Name of each cyanobacterium is further linked to a table that lists all the proteins

coded by its genome, along with their protein ID, locus ID, gene name and function. Upon a single click on any protein, an access page appears that displays links to physico-chemical properties, amino acid composition, biochemical pathway and structure information of the selected protein (**Figure 9**). User can select more than one protein from the protein-list of a cyanobacterium and export the data in CSV format. In addition, the user can also download the secondary structure, protein sequence and amino acid composition of the selected proteins.

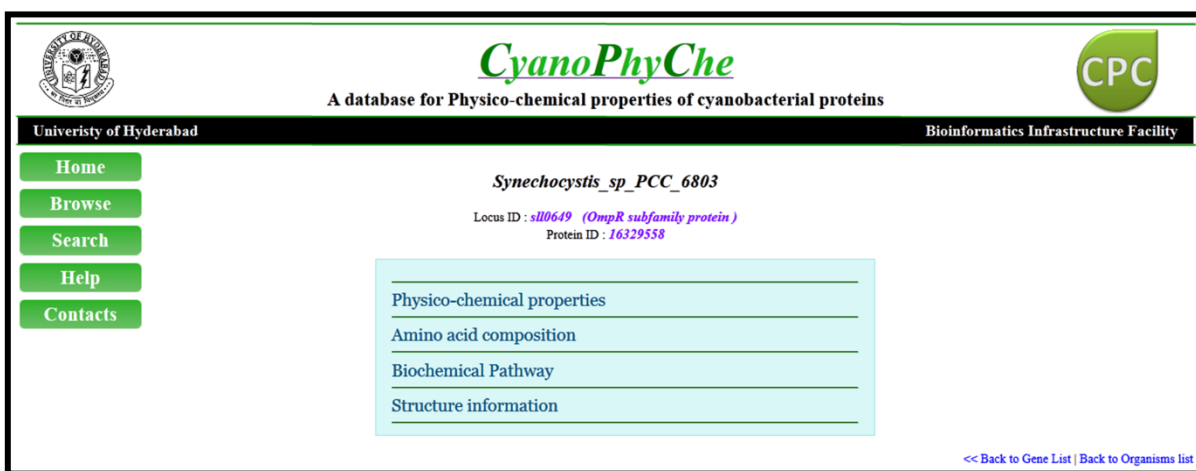


Figure 9: Snapshot of an access page of CyanoPhyChe. This page contains links to physico-chemical properties, amino acid composition, biochemical pathway and structure information.

2.2.3 Physico-chemical properties

Browsing through the link ‘Physico-chemical properties’ leads to a table containing physico-chemical properties of the protein in question. In addition, four different property scales with gradient-colored arrows are provided to aid the user to predict the solubility, polarity and structural stability of selected protein (**Figure 10**). Scales are provided for the canonical variable (CVsol), probability of an expressed protein entering into an inclusion body (PEPIB), GRAVY and structural stability (**Figure 10**). When the user browses through “Physico-chemical properties” of a protein for visualizing its properties, canonical variable for

solubility (CVsol), GRAVY value, PEPiB value, and structural stability are indicated on their respective scales (**Figure 10**). Property values of a selected protein are highlighted with blinking on the respective scales. **Figure 10** shows the determined physico-chemical properties of three selected proteins, Sll0649, Sll0088, and Sll0359 from *Synechocystis* sp. PCC 6803. PEPiB values calculated for Sll0649 and Sll0088 are 0.7 and 0.65, respectively. These values indicate that there is a high chance for these proteins to enter into inclusion body during their heterologous expression. To validate our predictions, we expressed these proteins to verify whether the calculated properties of the proteins can be relied upon and be considered for optimizing the conditions of expression, purification and characterization. Upon expression, most of the Sll0649 and Sll0088 were found in the inclusion body and a little was observed in the soluble fraction (Figure 11A and 11B). PEPiB value of Sll0359 protein is 0.1, hence it is predicted to be a soluble protein during its heterologous expression. As predicted, this protein was appeared in soluble fraction upon expression in *E. coli* (Figure 11C). These results are well in agreement with the calculated properties. The molar extinction coefficients calculated for the above proteins are 24750, 27310 and 7680 M⁻¹ cm⁻¹, respectively. These values can be used to determine the concentration of purified proteins by measuring their absorbance at 280 nm. Temperature is one of the factors that affect the structure of a protein. Studies show that there is a positive correlation between the structural stability and aliphatic amino acid content of proteins (Patrick Argos 1979; Ikai, 1980). The aliphatic index values of Sll0649, Sll0088 and Sll0359 are 95.1, 87.0 and 75.5. These values give the structural stability value of 4.8, 4.4 and 3.8 for these three proteins, respectively. This indicates that these proteins are moderately stable. The isoelectric point of a protein is an important property, because protein is least soluble at the pH near this point. There is a significant relation between pI of a protein and pH of the buffer being

used for crystallization. The buffer pH equals to or very near to the pI value of a protein a reasonable probability of yielding crystals (Kantardjieff and Rupp, 2004). The CyanoPhyChe

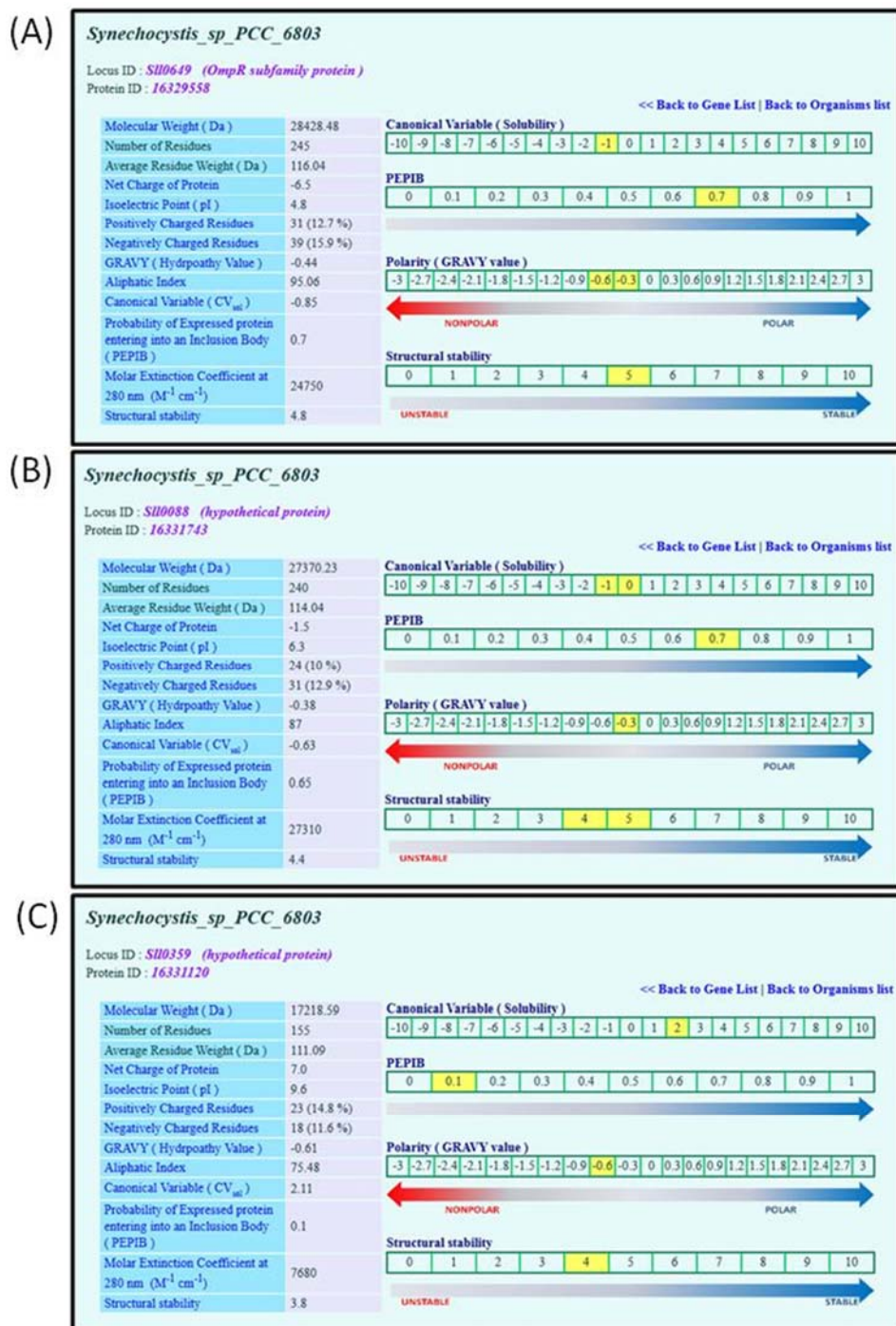


Figure 10: Snapshot of physico-chemical properties of the selected proteins. Physico-chemical properties of (A) SII0649, (B) SII0088 and (C) SII0359 proteins from *Synechocystis* sp. PCC6803. Based on the presented physico-chemical properties of solubility, probability of the protein entering into an inclusion body (PEPiB), polarity and structural stability were calculated and indicated on a point scale for better visualization of its nature.

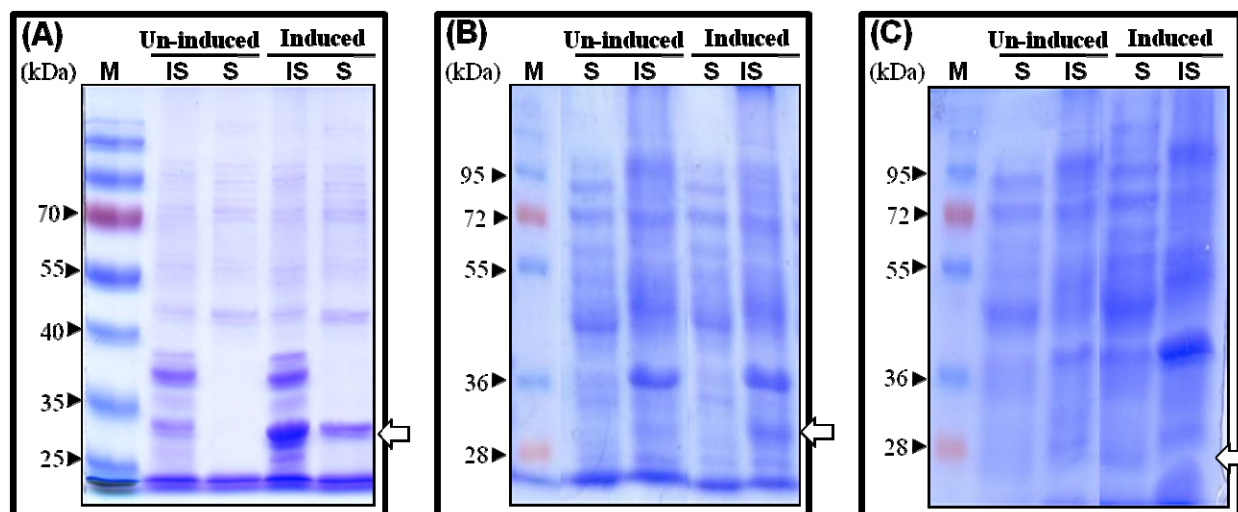


Figure 11: SDS-PAGE analyses of soluble and insoluble fractions of *E. coli* expressing the *Synechocystis* sp. PCC6803 proteins. Solubility of (A) Sll0649, (B) Sll0088 and (C) Sll0359 proteins. 0.4 mM IPTG (final concentration) was added to *E. coli* cells for inducing expression. The cells were harvested for separation of soluble and insoluble protein fractions, 2 hours after induced expression as described above. IS, Insoluble fraction; S, soluble fraction. Expressed protein bands are shown by open arrows.

database provides pI value for all the cyanobacterial proteins. Calculated pI of a cyanobacterial protein can be used to select a suitable buffer condition for its crystallization.

2.2.4 Amino acid composition and structure information

Amino acid composition and structural information of a cyanobacterial protein in question can be visualized by browsing through the links ‘Amino acid composition’ and ‘Structure information’, respectively. The ‘Structure information’ link navigates to the page containing predicted-secondary structure. In addition, an external link is provided to PDB database (<http://www.rcsb.org/pdb/home/home.do>) for viewing the 3D structure, when it is available. If the structure is not available for a protein, then the user can look for homologous proteins whose structures are available in the PDB database, by clicking on the “BLAST” tab for building a homology model.

2.2.5 Biochemical pathway in which a protein is involved

If a protein is known to be involved in any biochemical pathway, an external link to KEGG database is provided under the pathway name and ID that navigates to and displays the pathway in which it is involved (<http://www.genome.jp/kegg/pathway.html>).

2.2.6 Search options in the CyanoPhyChe database

Search page contains the list of cyanobacteria. User can choose one or more organisms from the list for finding proteins with desired properties. The selection leads to a query page, to retrieve a protein or a group of proteins with specific properties from the selected cyanobacteria (**Figure 12**). User can search the database using identifiers, such as gene name, locus ID, E.C. number, function and protein ID. The search window allows the user to enter a string of values for the selected identifier to search for different proteins among the selected organisms. The user can also search for the proteins based on the properties like molecular weight, number of residues and pI value. An option is provided to search and display proteins of a cyanobacterium that fall within a given range of a property. For instance, searching the database to retrieve the proteins with pI values ranging from 8 to 9 in *Synechocystis*, displays the list of proteins fall within this range. User can also retrieve the cyanobacterial protein(s) by searching the database using a specific property value or combination of two or more properties listed under “Search with combination of properties” menu. This option is more useful for retrieving proteins with different combination of properties, such as molecular weight, number of residues and isoelectric point. Physicochemical properties of the listed proteins from a single or multiple cyanobacterial species, using above search criteria can be exported in the CSV format. Additionally, the user can also export the secondary structure, protein sequence and amino acid

composition of the resulted proteins. These features of the database will be more useful to the researchers working on proteome analysis of any cyanobacterium.

Figure 12: Snapshot showing the CyanoPhyChe ‘Search’ page. A dropdown menu provided in the search page with various parameters can be used for retrieving cyanobacterial proteins in the database. Proteins of a cyanobacterium can be retrieved by gene name, locus ID, E.C. number and protein ID, and properties like molecular weight, number of residues, and pI value. User can also retrieve the cyanobacterial protein(s) by searching the database using combination of two or more properties listed under “Search with combination of properties” menu.

2.3 Conclusion

In summary, the database CyanoPhyChe is a collection of the calculated physico-chemical properties, solubility, and probability of an expressed protein entering into an inclusion body, structural stability, polarity and secondary structure of all cyanobacterial proteins. External links to PDB structure and KEGG pathway are provided in the database. Search option facilitates the retrieval of proteins of a particular cyanobacterium with specific property or combination of more than one property. The database also allows the user to export the retrieved data and encourages to use it for comparative studies. The data provided in the database can be used by the researchers, who are working on the cyanobacterial proteins for optimizing the methods employed for expression, purification, and characterization. The database is also useful

for interpreting the results obtained from proteome analysis of cyanobacteria. CyanoPhyChe will be further updated with additional information on cellular localization of cyanobacterial proteins and physico-chemical properties of the proteins encoded by the plasmid-DNA of cyanobacteria in the upcoming versions. Further, the database will be constantly updated and curated by the authors, as and when new information is reported in the literature or communicated by the users.

Chapter 3

Objective 2

ProtPhyChe: A automated web server for *in silico* characterization of prokaryotic proteins.

ProtPhyChe: A automated web server for *in silico* characterization of prokaryotic proteins

Summary

We developed 'ProtPhyChe' web server, which calculates physico-chemical properties, predicts secondary structure and builds homology models of prokaryotic proteins. All these tasks will be automatically performed for protein(s) upon either selecting them from the list of prokaryotic proteomes provided in the server or upon submission of the protein(s) of user's choice. ProtPhyChe works in two modes i.e. 'Normal' mode and 'Comparative' mode. In normal mode, physico-chemical properties, secondary and tertiary structures can be generated for selected proteins. In comparative mode the physico-chemical properties and tertiary structures can be compared among two or more orthologous proteins from any prokaryotic organisms listed in the web server. We have included a total of 1840 proteomes of sequenced prokaryotic organisms in the server. The user can select any prokaryotic organism of interest for its whole proteome *in silico* characterization or can also chose the protein(s) of interest to characterize them. We have also included an option to upload the newly sequenced proteome files for *in silico* characterization. Using ProtPhyChe, by comparison of the proteins and/or proteomes of two or more different microorganisms of interest, researchers get insights into their adaptation, evolution and ecology. ProtPhyChe can be accessed from <http://jssplab.uohyd.ac.in/pcp/index/index.php>

3. Introduction

During the past decade many number of prokaryotic genomes were sequenced. Availability of these sequenced genomes and their corresponding proteomes in public database provide a valuable raw resource for the researchers working in different fields of genomics and proteomics (Loman et al., 2012; Stahl and Lundeberg, 2012). By functional and comparative genome analysis of the sequenced microbial genomes, researchers get insights into their adaptation, evolution and ecology (Merhej et al., 2009; Gan et al., 2013; Gao et al., 2014). For structural and functional characterization of a protein various biophysical and biochemical methods are applied. Apart from experimental methods, different mathematical formulae were derived to determine physico-chemical properties of proteins (Ikai, 1980; Kyte and Doolittle, 1982; Harrison, 2000). Using these mathematical and computational tools, an unknown protein can be characterized *in silico* (Ashokan and Pillai, 2008; Bhattacharjee et al., 2008; Smith and Plazas, 2011; Pradeep et al., 2012). Several computational tools were developed for *in silico* prediction of physico-chemical properties, secondary structure elements, and to build 3D structure of a given protein based on homology modeling (Sali and Blundell, 1993; McGuffin et al., 2000; Rice et al., 2000). It is well known that the physico-chemical properties of protein, such as length, hydropathy, composition and properties of amino acid residues, and iso-electric point influence its stability and its folding (Levene PA, 1923; Chou and Fasman, 1978; Ikai, 1980; Kyte and Doolittle, 1982; Gill and von Hippel, 1989; Wilkinson and Harrison, 1991; Nakashima and Nishikawa, 1994; Idicula-Thomas and Balaji, 2005). Using *in silico* determined properties, probability of a protein entering into inclusion bodies (PEPIB) was predicted and the predictions were experimentally verified for selected proteins (Harrison, 2000). The studies involved in comparison of the physico-chemical properties of proteins from closely

related organisms showed that there exists the relation between the physico-chemical properties such as iso electric point, length of proteins, taxonomy and ecology of organisms (Kiraga et al., 2007). Comparison of secondary structure elements such as helices, coils and sheets of conserved proteins would give researcher an idea about the structurally conserved regions of a protein in question (Sitbon and Pietrokovski, 2007). Similarly the studies such as comparison of protein structures, docking, molecular simulations of the modeled structures would give insights into their structural stability, adaptability, and also elucidate the function of unknown protein (Szilagyi et al., 2002; Kumwenda et al., 2013; Maharaj and Soliman, 2013). Hence, comparison of physico-chemical properties, prediction of secondary structure and tertiary structures of proteins among prokaryotic organisms seems to be very important to gain insights into their evolution and adaptation to environmental niche. To aid the researchers in determining all the above said three properties in one go, we developed a web server 'ProtPhyChe'. By the use of ProtPhyChe any prokaryotic protein(s) can be characterized *in silico* in one go. ProtPhyChe, is capable of predicting the physico-chemical properties of whole proteome of selected micro organism, and it is also capable in generation of secondary and tertiary structure of the proteome or protein(s) in question. An additional feature of ProtPhyChe is that it capable of generating orthologs and generate physico-chemical properties, secondary structure and tertiary structure for the generated orthologs. This feature aids the researcher to have cross comparative study between the closely related organisms.

3.1 Materials and Methods

3.1.1 Construction of ProtPhyChe

The front end of the ProtPhyChe was developed using HTML, CSS, AJAX, jquery, PHP, and MySQL. All the programs running at the back end of the server, for analysing the user submitted data were written in Perl.

3.1.2 User registration

We have created sessions for the users of ProtPhyChe. The user has to first register with a user name and a password to use the server. The process of registration, login, and logout options were created using PHP and MySQL. Upon submitting the registration form, ProtPhyChe creates a unique ID for the user. All the generated data for the user submitted job will be stored in the directory name with this unique ID.

3.1.3 Selection of target organism

We have designed PHP scripts combined with HTML forms for selecting a target organism or to upload a file containing protein sequence(s) of user's choice. We have incorporated all protein sequence files (*.faa file) of 1840 micro-organisms, which were obtained from NCBI (<ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria>). List of the organisms are displayed in the web server. Upon selection or by uploading a protein sequence file, ProtPhyChe prompts the user for mode selection.

3.1.4 Normal mode and Comparative mode of *in silico* characterization

In normal mode, determination physico-chemical properties, prediction of secondary structure and tertiary structure of proteins will be carried out. The methodology used in prediction of physico-chemical properties in ProtPhyChe is similar to that CyanoPhyChe

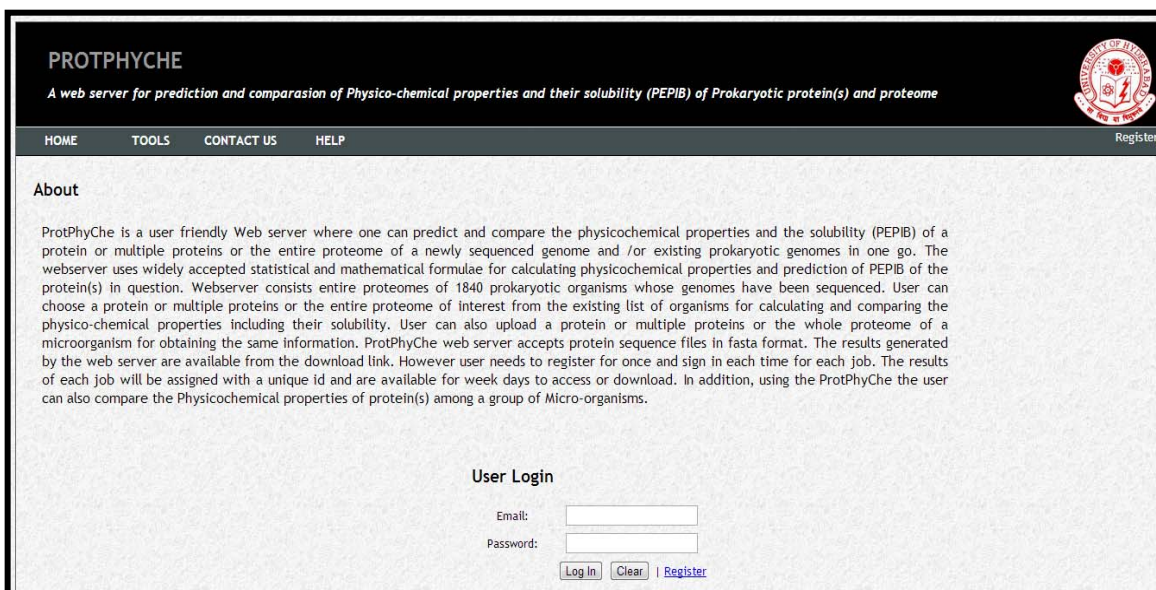
(Arun et al., 2012). For the prediction of aliphatic index, GRAVY and PEPPIB values for the proteins, we used Perl programs (with modifications in the code) those were used in CyanoPhyChe. These Perl programs use the mathematical equations proposed earlier for the prediction of aliphatic index, GRAVY and PEPPIB (Ikai, 1980; Kyte and Doolittle, 1982; Wilkinson and Harrison, 1991). We have integrated 'Predator' for the prediction of secondary structure elements of proteins (Frishman and Argos, 1997). For the prediction of tertiary structure, we integrated Modeller (Sali and Blundell, 1993) with ProtPhyChe. Initially for the selected target organism or uploaded file, the proteins present in them are given as input to PEPSTATS (primary seed data). The output obtained from PEPSTATS is taken as the secondary seed data and given for the prediction of aliphatic index, GRAVY and PEPPIB. The primary seed data is also given to 'Predator' and 'Modeller' for prediction secondary and tertiary structure. A Perl script was developed to retrieve the protein sequences from the PDB files and create the local database. During the time of prediction of tertiary structure, a couple of Perl programs are invoked for performing BlastP between the query protein and the local protein database build out of protein sequences retrieved from PDB files. Thereafter, the Perl programs retrieve the suitable template which meet the prescribed percentage identity with the query protein for building homology model by 'Modeller'. In comparative mode, both target and reference organisms are to be selected by the user for identification of orthologs by using bidirectional best hit method. The orthologous protein sequences of selected target and reference organisms will be retrieved and then set for prediction of physico-chemical properties, secondary structure and tertiary structure as in normal mode. Upon completion of whole process either in normal mode or comparative mode, the results will be displayed in the web browser. This task was achieved by developing a PHP script, which displays the results in tabular format in a web

browser. It was reported that at the time of homology modelling the identity between the query protein and the template should be greater than or equal to 30% (Schwede et al., 2003). We developed Perl scripts, which perform both local and global alignment between query and template protein sequences. For performing the global and local alignments between the query protein and template protein sequences we used standalone version of 'Needle' and 'Water' tools from EMBOSS package.

3.2 Results and Discussion

3.2.1 Description of ProtPhyChe

The web interface contains 'Home', 'Tools', 'Contact', 'Help'. A brief introduction about ProtPhyChe is given in the home page (**Figure 13**). For performing the tasks of determination of physico-chemical properties, secondary and tertiary structure predictions, the user needs to login into the web server.



PROTPHYCHE
A web server for prediction and comparison of Physico-chemical properties and their solubility (PEPIB) of Prokaryotic protein(s) and proteome

HOME TOOLS CONTACT US HELP Register

About

ProtPhyChe is a user friendly Web server where one can predict and compare the physicochemical properties and the solubility (PEPIB) of a protein or multiple proteins or the entire proteome of a newly sequenced genome and /or existing prokaryotic genomes in one go. The webserver uses widely accepted statistical and mathematical formulae for calculating physicochemical properties and prediction of PEPIB of the protein(s) in question. Webserver consists entire proteomes of 1840 prokaryotic organisms whose genomes have been sequenced. User can choose a protein or multiple proteins or the entire proteome of interest from the existing list of organisms for calculating and comparing the physico-chemical properties including their solubility. User can also upload a protein or multiple proteins or the whole proteome of a microorganism for obtaining the same information. ProtPhyChe web server accepts protein sequence files in fasta format. The results generated by the web server are available from the download link. However user needs to register for once and sign in each time for each job. The results of each job will be assigned with a unique id and are available for week days to access or download. In addition, using the ProtPhyChe the user can also compare the Physicochemical properties of protein(s) among a group of Micro-organisms.

User Login

Email:

Password:

Figure 13: Snapshot showing the home page of ProtPhyChe web server. The user can register from the registration page obtained by clicking on the 'Register' option provided at the top right end of the screen. With the registered user credentials the email and password can be entered into the text area for login.

Login credentials can be obtained by clicking on the 'Register' link. Upon a single click on the 'Register' link, the home page navigates to registration page. The registration page contains the fields such as 'Name', 'E-mail ID', 'Password', 'Confirm Password' and 'Password Hint'. User has to enter a name, valid e-mail address and password in the fields provided. These entries are used as login credentials by the user to start ProtPhyChe. Just below the registration there is CAPTCHA. The user has to enter the characters displayed in the image in the text box provided for security purpose of the server. Below the random text field there are two buttons labeled as "Reset" to re-enter the entire information and "Go" button for submitting registration form.

3.2.2 Starting ProtPhyChe

After successful registration, the registration page automatically redirects to home page for login. Once the user enters the registered email address and password into the fields provided, the login page navigates to a new page as shown in **(Figure 14)**.

3.2.3 Selection of a proteins or proteome

A selected prokaryotic organism, whose protein(s) are to be characterized *in silico* is here after referred as target organism. As shown in **Figure 14**, user can select any micro organism as target from the list. Below the organisms list there is a '*Search proteins*' tab. We can select desired protein(s), by typing keyword(s), from the entire proteome of the target organism. The keywords entered are not case sensitive. By default "*None*" is present in the text box, which means that the server will predict physico-chemical properties, secondary and tertiary structures for all the proteins of a selected target organism upon clicking on 'GO' button.

3.2.4 Uploading file containing protein sequences

Beside the selection of proteome from the list of organisms displayed, the user can also upload a file containing protein sequences of interest (**Figure 14**).

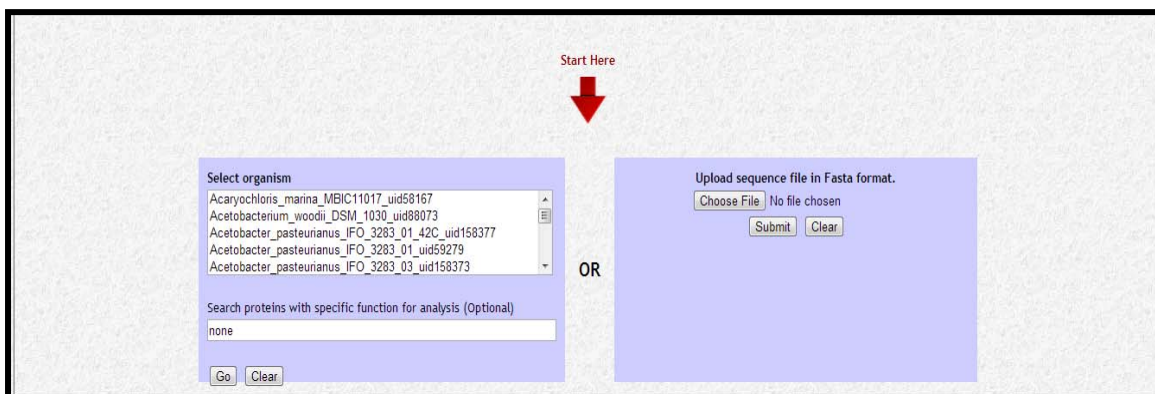
The screenshot shows a web interface with two main options separated by an "OR" label. On the left, under "Select organism", there is a scrollable list of organisms including Acaryochloris_marina_MBIC11017, Acetobacterium_woodii_DSM_1030, and Acetobacter_pasteurianus with various IFD and GID IDs. Below this list is a search box labeled "Search proteins with specific function for analysis (Optional)" with the text "none" and "Go" and "Clear" buttons. On the right, under "Upload sequence file in Fasta format.", there is a "Choose File" button, a status "No file chosen", and "Submit" and "Clear" buttons. A red arrow labeled "Start Here" points to the top of the interface.

Figure 14: Snapshot showing the options provided in the ProtPhyChe for selection of an organism or to upload file. User can select any organism from the list displayed. There is a text box provided for the user to filter the specific user interested proteins from the whole proteome. The user can provide any keywords such as ORF ID, GID, protein function etc for filtration. If there is a need to work on more than one protein then the keywords are given as string with a comma between each keyword. The user also has an option to upload a fasta file containing one or more than one protein sequences of his own.

3.2.5 Working modes of ProtPhyChe

3.2.5.1 Normal Mode

In normal working mode of ProtPhyChe, the Perl programs which run at the backend of the ProtPhyChe perform the tasks such as prediction of physico-chemical properties, prediction of secondary and tertiary structure of the protein(s). During the time of analysis the user can logout from the link provided below the header file. To view the status of a submitted job, the user needs to login again. After completion of the job, all individual results for each protein can be viewed in the web page in a tabular form and the same data is also available for download in ZIP format. In addition to the physicochemical properties, the web server also provides PEPID values, which is an indicator of whether a protein enters into an inclusion bodies or not upon heterologous expression (Arun et al., 2012). Another important

feature of ProtPhyche is that user can download proteins with in a pI range of a selected prokaryote. This feature would aid researchers to cross verify and get an idea on the protein spots that appear on 2D gel electrophoresis in a given pH range. Thus, the Protphyche is very useful to the researchers performing proteome analysis of microorganisms.

3.2.5.2 Comparative Mode

In comparative mode, list of proteomes are displayed as shown in **Figure**

15.

Your selections are

Organism : *Acaryochloris_marina_MBIC11017_uid58167*

Selected Keyword : none

To proceed with above choices select the working mode and submit the job or [reselect](#) the organism.

Your file ID : *FX2123E1427*

Select Working mode proceed further:

☐ Normal (For Prediction of Physico-Chemical properties)

☒ Comparative (For Prediction of Physico-Chemical properties and comparason among orthologs)

Select organism for comparative analysis:

- ☐ *Acaryochloris_marina_MBIC11017_uid58167*
- ☐ *Acetobacter_lun_woodii_DSM_1030_uid88073*
- ☐ *Acetobacter_pasteurianus_IFO_3283_01_42C_uid158377*
- ☐ *Acetobacter_pasteurianus_IFO_3283_01_uid59279*
- ☐ *Acetobacter_pasteurianus_IFO_3283_03_uid158373*
- ☐ *Acetobacter_pasteurianus_IFO_3283_07_uid158381*
- ☐ *Acetobacter_pasteurianus_IFO_3283_12_uid158379*
- ☐ *Acetobacter_pasteurianus_IFO_3283_22_uid158383*
- ☐ *Acetobacter_pasteurianus_IFO_3283_26_uid158531*

E-Value Note: E-value should be "d" (BLASTP default e value) or in form of "1e-10".

% alignment

Figure 15: Snapshot describing the options of comparative mode. When the user selects the radio button representing “Comparative” indicates that the user has set the working style of the server as comparative mode. When the comparative mode is selected by the user, then automatically a drop down menu displaying the list of bacterial genomes are displayed. From this list of genomes the user can select the closest relatives of the target genome or for uploaded file, as the reference organisms. The server predicts the bidirectional best hits, and then calculates physico-chemical properties, secondary structure and tertiary structure of the proteins orthologs proteins along with the proteins in question.

The user can select any organism or multiple organisms as reference organisms. Upon clicking on 'Proceed' button, the bidirectional best hits are generated for the proteins of target organism and the rest of the process takes place as in normal mode.

3.2.6 Format of the results:

The predicted results of both normal and comparative modes can be viewed by clicking on the "view" option available on the server activity page (**Figure 16**).

S No.	Job ID	Status	Mode	Date	Time	Results	Link
1	AF7889M7632	completed	Comparative	2014-11-26	14:41:36	Refresh	View Download
2	DQ2337N1255	completed	Comparative	2014-11-26	14:18:00	Refresh	View Download
3	SZ6968Y3116	completed	Comparative	2014-11-26	14:12:04	Refresh	View Download
4	YJ9057R6901	completed	Comparative	2014-11-26	14:08:22	Refresh	View Download
5	HS1832V6377	completed	Comparative	2014-11-26	13:55:35	Refresh	View Download
6	LH7552N6830	completed	Comparative	2014-11-26	13:13:07	Refresh	View Download
7	NO8772I5789	completed	Comparative	2014-11-26	12:31:35	Refresh	View Download
8	RS5923T9339	completed	Normal	2014-11-26	11:56:16	Refresh	View Download
9	RY2060T1187	completed	Normal	2014-11-26	11:54:14	Refresh	View Download
10	WZ5815D1024	completed	Normal	2014-11-08	10:10:41	Refresh	View Download
11	ZE8974A9922	completed	Comparative	2014-11-08	10:03:05	Refresh	View Download
12	KM1001G1090	completed	Normal	2014-10-27	17:55:15	Refresh	View Download

Figure 16: Snapshot showing the server activity of ProtPhyChe. User can view the Job ID of the present job, status of the job whether the submitted job is running or completed, Mode selected for the job, date of submitting the job, time of submitting job, link to refresh the status of the job, view link for viewing the predicted results in web browser and download link for downloading the results. Download link would be appearing automatically when the status of the job is shown as completed.

In normal mode the predicted physico-chemical properties are displayed in a table and a download link is available on the bottom of the screen. In comparative mode the results are displayed similar to that of normal mode in the form of a table. The ORF numbers of target and reference organisms are further hyperlinked to the corresponding protein's physico-chemical properties data. The downloaded ZIP file contains the physico-chemical properties, secondary structure and tertiary structures data.

3.3 Applications of ProtPhyChe

3.3.1 Normal mode application

We have selected '*Synechocystis* PCC 6803 substr GT I' as the target organism and submitted it for whole proteome *in silico* characterization in normal mode to test the performance of ProtPhyChe. As described above, in normal mode, physico-chemical properties, secondary structure and tertiary structures were generated for *Synechocystis* by the web server. **Figure 17** shows the partial list of predicted physico-chemical properties of 3169 proteins of *Synechocystis*.

ORF-No	Function	PEPIB	Molecular Weight	Total Residues	Average residue weight	Charge	Isoelectric Point	A280 Molar Extinction Coefficients	Aliphatic Index	GRAVY	Tiny amino acids(A,C,G,S,T and %)	Small amino acids(A,C,D,G,N,P,S,T,V and %)	Aliphatic amino acids(A,I,L,V and %)	Aromatic amino acids(F,H,W,Y and %)
SYNPCCN_0001	solaneyl diphosphate synthase	0.8	35725.68	323	110.6	-12.5	4.7	15930	100.3	-0.0	89(27.6)	157(48.6)	125(38.7)	29(9.0)
SYNPCCN_0002	hypothetical protein	0.2	21028.94	185	113.7	10.5	10.8	31970	66.3	-0.7	53(28.6)	89(48.1)	44(23.8)	18(9.7)
SYNPCCN_0003	hypothetical protein	0.2	18297.06	173	105.8	0.5	7.0	22920	86.1	0.2	62(35.8)	93(53.8)	61(35.3)	20(11.6)
SYNPCCN_0004	hypothetical protein	0.4	26465.02	233	113.6	-0.5	6.3	70930	81.1	-0.4	64(27.5)	105(45.1)	68(29.2)	31(13.3)
SYNPCCN_0005	magnesium-protoporphyrin IX monomethyl ester cyclase	0.4	42153.21	358	117.7	2.5	7.1	55350	70.4	-0.4	79(22.1)	152(42.5)	93(26.0)	59(16.5)
SYNPCCN_0006	hypothetical protein	0.5	34719.58	312	111.3	-3.0	5.7	39880	78.6	-0.3	85(27.2)	152(48.7)	95(30.4)	34(10.9)
SYNPCCN_0007	GDP-D-mannose dehydratase	0.5	41333.80	362	114.2	-0.5	6.4	57300	77.7	-0.4	96(26.5)	170(47.0)	104(28.7)	49(13.5)
SYNPCCN_0008	photosystem II D1 protein	0.3	39721.40	360	110.3	-5.0	5.5	75860	80.8	0.3	118(32.8)	187(51.9)	113(31.4)	62(17.2)
SYNPCCN_0009	arginine decarboxylase	0.6	74476.86	659	113.0	-14.0	5.2	84690	90.0	-0.3	166(25.2)	299(45.4)	208(31.6)	73(11.1)
SYNPCCN_0010	NAD-dependent DNA ligase Liga	0.5	74602.23	669	111.5	-8.5	5.3	70820	91.3	-0.3	173(25.9)	315(47.1)	221(33.0)	53(7.9)
SYNPCCN_0011	hypothetical protein	0.5	23093.24	202	114.3	-5.0	5.0	29910	76.6	-0.3	56(27.7)	83(41.1)	52(25.7)	28(13.9)
SYNPCCN_0012	iron(III) dicitrate transport system permeaseprotein	0.2	35693.46	343	104.1	7.0	9.8	44920	121.1	1.0	126(36.7)	188(54.8)	159(46.4)	32(9.3)
SYNPCCN_0013	iron(III) dicitrate transport system permeaseprotein	0.2	36770.03	349	105.4	7.5	8.7	49960	127.2	1.0	117(33.5)	187(53.6)	164(47.0)	31(8.9)

Figure 17: Snapshot showing the partial list of predicted physico-chemical properties of *Synechocystis* proteome in normal mode of ProPhyChe.

Full set of physico-chemical properties of entire proteome of *Synechocystis* generated by the web server are downloaded for validation. The secondary structures predicted for all the proteins of the target organism are obtained in a folder named 'secondary_structure' in the downloaded ZIP file. **Figure 18** shows the predicted secondary structure for the protein Sds (Solanesyl diphosphate synthase) encoded by the ORF *synpccn_0001*.

[illegible]

The predicted tertiary structure of the proteins can be viewed in protein structure visualization tools such as Pymol (The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC.). **Figure 19** shows the predicted tertiary structure for 'Solanesyl diphosphate synthase' encoded by the ORF *synpccn_0001* of *Synechocystis*.

3.3.2 Comparative mode application

In comparative mode of analysis, the 'Sds' protein from '*Synechocystis* PCC 6803 substr PCC P' (target organism) and '*Thermosynechococcus elongates* BP 1' (reference organism) was chosen. *Synechocystis* is a mesophile and lives in fresh water, where as '*Thermosynechococcus elongatus* BP 1' is a thermophile and is found in hot springs. We have selected Sds protein to compare the protein properties and structure between these two organisms, because they thrive in different habitats.



Figure 19: Snapshot showing the predicted tertiary structure for the protein 'Solanesyl diphosphate synthase' encoded by *synpccn_0001* of *Synechocystis* PCC 6803 substr GT I.

Hence, comparison of proteins from these cyanobacterial species, adapted to survive under extremely different environmental niche would give insights into the protein (molecular) evolution with respect to the organism's adaptation to different habitat. **Figure 20** shows physico-chemical properties of 'Solanesyl diphosphate synthase' (Sds)' proteins of *Synechocystis* and *Thermosynechococcus*.

Genome	Function	Protein-ID	PEPIB	Molecular weight	Residues	Average Residue Weight	Charge	Isoelectric Point	A280 Molar Extinction Coefficients	Tiny amino acids(A,C,G,S,T)
Thermosy solanesyl		22299300	0.71	35336.26	323	109.4	-10	4.6948	20400	100 (31.0)
Synechoy solanesyl		383320910	0.81	35725.68	323	110.606	-12.5	4.7105	15930	89 (27.6)

Figure 20: Snapshot showing the predicted physico-chemical properties of protein 'Solanesyl diphosphate synthase' of *Synechocystis* PCC 6803 substr PCC P and its ortholog from *Thermosynechococcus elongatus* BP-1. We can observe the differences in charge, iso-electric point etc.

Homology models built for orthologs from different bacteria by the web server (as described in materials and methods) can be compared using PyMoL. The Sds homology models of both *Synechocystis* and *Thermosynechococcus*, which were built in comparative mode by the web server were super-imposed in PyMoL to examine the structural variations. **Figure 21** shows the super imposed structures of Sds protein of *Synechocystis* sp. PCC 6803 substr PCC P (green colour) and *Thermosynechococcus elongatus* BP 1 (red colour). It is clear from the super

imposed structures that there is a bend in the helices and difference in the loop formation. Similarly one can compare the secondary structure of Sds of ortholog of *Synechocystis* with that of *Thermosynechococcus*. In this way by analyzing the results obtained from comparative mode, one can clearly perform the comparative analysis of physico-chemical properties, secondary and tertiary structures of orthologs from various bacteria, provides clues for better understanding the key factors responsible for the protein stability and their adaptation.



Figure 21: Snapshot showing the super imposed structures of protein 'Solanesyl diphosphate synthase of *Synechocystis* sp. PCC 6803 substr PCC P (green colour) and its ortholog from *Thermosynechococcus elongatus* BP-1 (red colour). From the super imposed structure, we can observe the clear cut difference in loop formation (circled in red)

3.4 Conclusion

The researcher can *in silico* characterize any prokaryotic protein in one go using ProtPhyChe, instead of using multiple tools. ProtPhyChe generates, physico-chemical properties, builds secondary and tertiary structures of selected proteins of target as well as reference organisms. By selecting a protein(s) of interest in a target organism and by choosing the reference organism, ProtPhyChe identifies the orthologs for *in silico* characterization in comparative mode. An important feature of ProtPhyChe is that any physico-chemical properties

and structural information of any prokaryotic protein can be compared with its orthologs for better understanding of molecular evolution, factors that influence function of a protein. Another advantage of ProtPhyChe is that the researcher need not obtain protein sequences or whole proteome from publicly available databases. The entire proteome data of sequenced microbial organisms has been integrated in the web server. However, an option to upload new protein sequence(s) to the web server is provided for analysis of newly sequenced microbial genomes. The results generated by the ProPhyChe can be viewed in MS-Excel. The predicted secondary structures can be viewed using any text editor. Tertiary structures can be viewed using PyMOL or RasMol visualization tools. ProtPhyChe provides preliminary idea on probability of an expressed protein entering into inclusion bodies (PEPIB). Such information can be used to optimize conditions for expressing proteins into soluble form. The Physico-chemical properties results obtained upon using comparative mode would help the researcher to understand the concepts of adaptation, niche differentiation, molecular evolution and ecology.

Chapter 4

Objective 3

CyanoCis: A web tool for identification of *cis*-acting elements in cyanobacterial genomes

CyanoCis: A web tool for identification of *cis*-regulatory elements in cyanobacterial genomes

Summary:

CyanoCis is a web-tool developed for identification of the *cis*-regulatory elements located in the upstreams to cyanobacterial gene(s). It is an automation of various steps of phylogenetic foot printing method and uses cyanobacterial genomes for finding *cis*-regulatory elements. The web tool first identifies orthologs for selected gene(s), generates clustered DNA upstreams of these orthologs and subsequently directs these clustered upstreams to MEME, Gibbs motif sampler, MD Scan and Bioprosector for motif prediction. As each motif prediction tool, integrated with CyanoCis, uses unique algorithm, a *cis*-regulatory element commonly generated by these tools, can be considered with confidence for further analysis. The CyanoCis is useful to researchers working on mechanisms of gene regulation and to build gene regulatory networks in cyanobacteria. CyanoCis can be accessed from http://jssplab.uohyd.ac.in/Cyanocis_2/

4. Introduction

Analysis of genome sequences of cyanobacteria, revealed the presence of several putative transcription factors and two-component regulatory systems consisting of a sensory histidine kinase and a cognate response regulator for control of gene expression (Kaneko et al., 1996; Paithoonrangsarid et al., 2004; Shoumskaya et al., 2005; Kanesaki et al., 2007; Vijayan et al., 2011). In addition to these transcriptional regulators, several non-coding RNAs were identified in different cyanobacterial species and the role of majority of these regulatory RNAs in regulation of gene expression needs further exploration (Axmann et al., 2005; Steglich et al., 2008; Voss et al., 2009). Although, regulons of several response regulators as well as transcription factors were identified in cyanobacteria, the genes regulated by many of them and their target DNA binding elements are yet to be identified (Los et al., 2008). Availability of cyanobacterial genome sequences and computational methods for the prediction of target DNA binding sites prompted us to generate a web server, for the identification of *cis*-regulatory elements located in the upstream of any given cyanobacterial gene(s). This would aid researchers for better understanding of regulation of gene expression and for building gene regulatory networks in cyanobacteria. We used automation of various steps involved in a widely accepted computational method, phylogenetic foot printing for prediction of conserved DNA elements (Ganley and Kobayashi, 2007).

4.1 Phylogenetic footprinting

The principle behind the phylogenetic foot printing is that, the functional elements of the genomes undergo slower rate of change than the non-functional elements (Ganley and Kobayashi, 2007). In phylogenetic foot printing, the non coding region of DNA, usually the

upstream sequences of the orthologous genes of closely related species are taken for analysis. Initially the orthologs for a protein in question will be identified among the closely related organisms, then the upstream sequences of these orthologous group are clustered and then used for the identification of *cis*-regulatory elements (Wels et al., 2006; Arun et al., 2013). The motif prediction tools such as MEME (Multiple Em for Motif Elicitation), Bioprosector, MD (Motif Discovery) and Gibbs-motif sampler can be used for the identification of *cis*-regulatory elements in the clustered upstream sequences of orthologous genes (Bailey and Elkan, 1994; Liu et al., 2001; Liu et al., 2002; Thompson W, 2007). We developed CyanoCis based on the phylogentic foot-printing and by integrating four motif prediction tools. CyanoCis is first of its kind and can be accessed from http://jssplab.uohyd.ac.in/Cyanocis_2.

4.2 Materials and Methods

4.2.1 Selection of cyanobacterial genomes

We have downloaded genome files of cyanobacteria with file extensions *.faa, *.fna and *.ptt from NCBI FTP link (<ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria>). We have considered the cyanobacterial species / strains having largest genome size among the multiple species / strains of the same genus, to avoid generation of false positive *cis*-regulatory elements by the CyanoCis web server. The cyanobacterial genomes used were *Synechocystis* sp. PCC 6803, *Acaryochloris marina* MBIC 11017, *Synechococcus* CC 9311, *Anabaena variabilis* ATCC 29413, *Synechococcus elongatus* PCC 6301, *Cyanothece* PCC 7424, *Synechococcus* JA-2 3B a 2 13, *Gloeobacter violaceus* PCC 7421, *Synechococcus* PCC 7002, *Microcystis aeruginosa* NIES 843, *Nostoc punctiforme* PCC 73102, *Thermosynechococcus elongatus* BP1, *Prochlorococcus marinus* MIT 9303, and *Trichodesmium erythraeum* IMS 101.

4.2.2 Identification of orthologs and generation of cyCoGs

We used Bi-directional best hit method (BDBH) to identify orthologs in other cyanobacteria, for each protein of a query cyanobacterium. The reciprocal BLASTP program with an E-value cutoff 10^{-3} in both directions was performed (Altschul et al., 1990). Each protein (query) of one cyanobacterial species (query organism) was searched against rest of the cyanobacterial proteins (reference organisms) and then best hit from each reference organism was searched against proteins of query organism. When the top and high scoring hits of both forward and reverse blast searches matches with each other, then it is considered as true orthologs. We did BDBH for fourteen times, each time one, out of 14 cyanobacterial species was taken as query and rest of them as reference organisms. We retrieved the top scoring hits and clustered them to generate cyanobacterial clusters of orthologs groups (cyCoGs). The cyCoGs containing a minimum of four orthologs or more were extracted for further analysis. Next, we generated cyanobacterial clusters of transcriptional units (cyCoTs) by reading the cyCoG files and the operon data obtained from DOOR database (Mao et al., 2009). If an ortholog of a cyCoG, is encoded by a structural gene, which is located downstream to the first gene of an operon, then the upstream DNA sequence of the first gene was used for motif prediction.

4.2.3 Clustering upstream DNA sequences of genes of a cyCoT

We extracted and clustered the upstream DNA sequences of cyCoTs by developing in house Perl programs. By reading the co-ordinates of each gene from the *.ptt file of a cyanobacterial genome, the intergenic distance between adjacent genes, as well as the position of this intergenic sequence in the genome were calculated. This information was used to retrieve intergenic (upstream DNA sequence of an ORF) sequence from *.fna file of respective cyanobacterial genome and saved in the form of text files. When the actual length of the

upstream sequence is less than that of 500 bp, we considered the entire intergenic DNA sequence as an upstream DNA sequence of a gene. Clustered upstream DNA sequences of all cyCoTs were generated, by grouping the upstream sequence of each gene of a cyCoT. Thus clustered upstream sequences of all cyCoTs of all selected cyanobacteria were generated and stored as back ground data. The list of selected cyanobacteria and their genes, cyCoGs, cyCoTs and clustered-upstream DNA sequences were used for constructing the CyanoCis web server. Upon selection of a gene(s) from a cyanobacterium by the user, the web server reads the stored-data in the back ground, displays the orthologs, and directs the clustered upstream DNA sequences of corresponding cyCoT(s) to motif prediction tools, which were integrated in CyanoCis, for *cis*-regulatory element prediction. We integrated four different motif prediction tools, each one of these work based on a unique algorithm. For generating clustered-upstream sequences of each cyCoT of all cyanobacterial genes, required prediction of *cis*-regulatory elements, have been generated using different Perl scripts. For running the integrated motif prediction tools we provided an HTML form, which contains default parameter values for prediction of *cis*-regulatory elements. The user either can change these parameters or use the default parameter values, such as width of the motif, number of motifs etc., as per the requirement. The CyanoCis web tool, automatically generates a file called “parameter file” upon submission of the above mentioned HTML form. Four specific Perl programs were designed for each motif prediction tool that are capable of processing the content of the parameter file and invoke the motif prediction tools to predict *cis*-regulatory elements. A PHP program was designed in such a way that, the predicted results are displayed on the screen and can also be downloaded in the form of a pdf file.

4.2.4 Parameters used for finding *cis*-regulatory elements for selected *Synechocystis* genes

We tested the performance of CyanoCis by selecting the genes *slr1756*, *sll0679*, *sll0822*, *smr0009*, *sll0622*, and *sll1916* from the gene list of the *Synechocystis*. After a single click on “Find *cis*-regulatory elements” the parameter page appeared. We changed width of the motif to be predicted to 25 for each motif prediction tool for identifying *cis*-regulatory elements.

4.3 Results and discussion

4.3.1 Description of CyanoCis

The web interface of the CyanoCis web server contains 'Home', 'Browse', 'Help', and 'Contact' links on the top left corner of the index page. Cyanobacterial genes can be selected by ‘Browse’ link. User can browse through the list of genes and can select any cyanobacterial gene(s) for prediction of *cis*-regulatory elements. A brief introduction to using CyanoCis is provided on the home page (**Figure 22**).

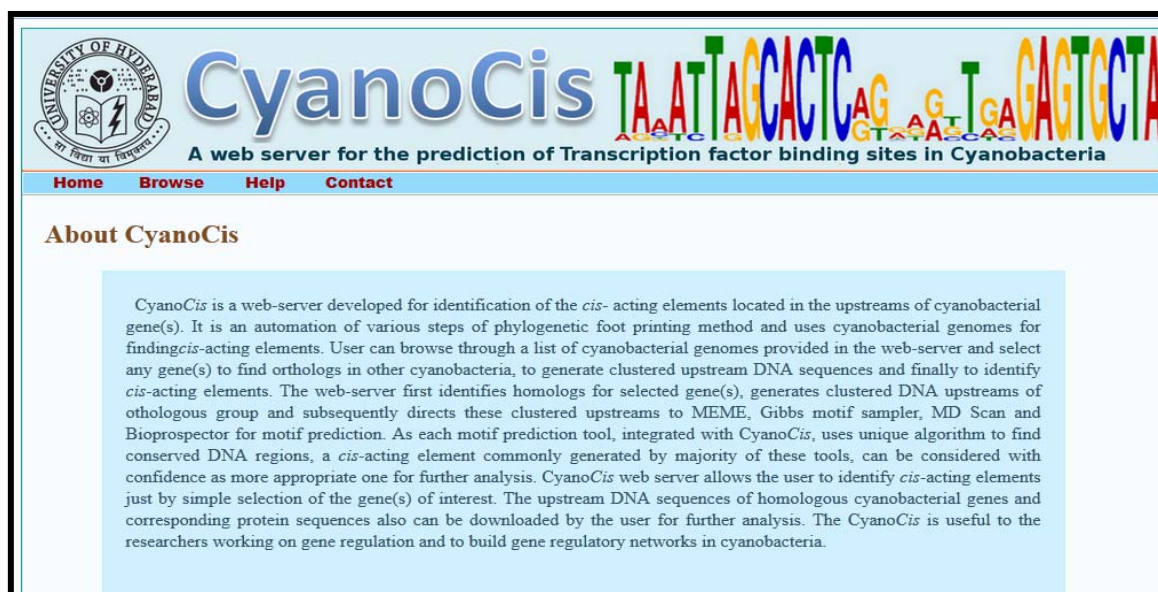


Figure 22 : Snapshot of the home page of CyanoCis web server. The home page contain 'Home', 'Browse', 'Help' and 'Contact' buttons. User can select cyanobacterial gene(s) of interest by browsing through ‘Browse’ link.

4.3.2 Interface of the CyanoCis

When the user navigates through the 'Browse' option, the list of fourteen cyanobacteria is displayed. Name of each cyanobacterium is further linked to a table that lists all the open reading frames (Locus ID) of its genome, along with gene name, protein ID and function (**Figure 23**). User can select any gene or multiple genes from the 'list of genes' of a

List of Genes

Synechocystis_PCC_6803

[<< Go Back to Organisms list](#)

S.No	Locus ID	Gene Name	Protein ID	Function
<input checked="" type="checkbox"/> 1	slr0612	-	16329172	hypothetical protein
<input checked="" type="checkbox"/> 2	slr0613	-	16329173	hypothetical protein
<input type="checkbox"/> 3	slr0558	-	16329174	hypothetical protein
<input type="checkbox"/> 4	slr1214	-	16329175	magnesium-protoporphyrin IX monomethyl ester cyclase
<input checked="" type="checkbox"/> 5	slr1213	-	16329176	hypothetical protein
<input type="checkbox"/> 6	slr1212	rfbD	16329177	GDP-D-mannose dehydratase
<input type="checkbox"/> 7	slr1311	psbA2	16329178	photosystem II D1 protein
<input checked="" type="checkbox"/> 8	slr1312	speA	16329179	arginine decarboxylase
<input type="checkbox"/> 9	slr1209	ligA	16329180	NAD-dependent DNA ligase LigA
<input type="checkbox"/> 10	slr1315	-	16329181	hypothetical protein
<input type="checkbox"/> 11	slr1316	fecC	16329182	iron(III) dicitrate ABC transporter permease
<input type="checkbox"/> 12	slr1317	fecD	16329183	iron(III) dicitrate ABC transporter permease
<input type="checkbox"/> 13	slr1318	fecE	16329184	iron(III) dicitrate ABC transporter permease
<input type="checkbox"/> 14	slr1319	fecB	16329185	iron(III) dicitrate ABC transporter permease

Figure 23: Snapshot showing the list of genes of cyanobacterium *Synechocystis* sp. PCC6803. The user can select any of the gene(s) from the list provided, using scroll button. A search tab is provided on right corner of the page for searching genes of interest using identifiers.

selected cyanobacterium. Alternatively, a search box is provided on top of the 'list of genes' for selecting a desired gene (s) by searching with identifiers. User can search using identifiers, such as gene name, locus ID, protein ID or function (**Figure 23**). After selection of gene(s) either using browse or search options, a single click on "submit" button navigates to a new page that displays the genes that are selected by the user, links to 'Retrieve Orthologs' and link to prediction of *Cis* -regulatory elements (**Figure 24**).

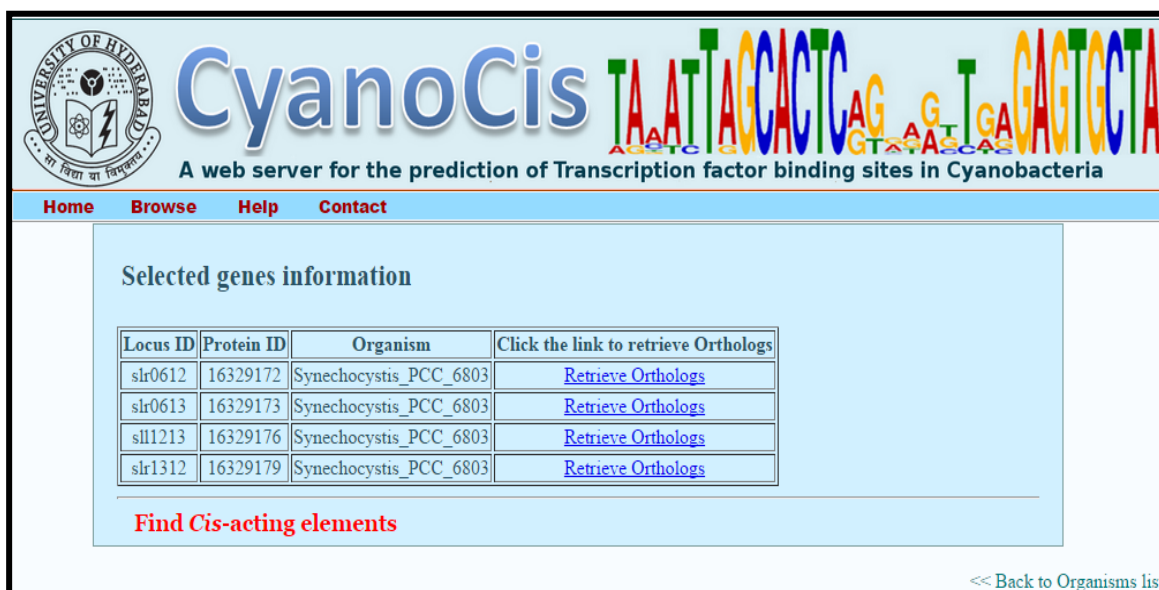


Figure 24: Snapshot showing the options for viewing the bidirectional best hits and prediction of *cis*-regulatory elements for the selected genes of *Synechocystis*.

4.3.3 Identification of orthologs of selected cyanobacterial genes

Upon a single click on "Retrieve orthologs" link will navigate to a new page showing a list of cyanobacteria with check boxes. The user can download orthologs of a gene or multiple genes, by selecting them using check boxes. Upon clicking "proceed" button, the orthologs of the selected gene are displayed. We provided a download link to download the orthologous protein sequences (**Figure 25 & 26**).

4.3.4 Prediction of *Cis* -regulatory elements

Cis-regulatory elements can be predicted for a single selected gene or multiple genes at a time. Upon selection of genes from the 'list of genes' of a cyanobacterium, a single click on 'predict *cis*-regulatory elements' tab leads to parameter page (**Figure 27**). Either the user can predict *cis*-regulatory elements with the default settings already provided or can change these parameters with the desired values.

Prediction of Orthologs for the organism: *Synechocystis_PCC_6803*

Gene: **slr0612**

Check the box for genome selection:

- ☒ Acaryochloris_marina_MBIC_11017
- ☒ Anabaena_variabilis_ATCC_29413
- ☒ Cyanothece_PCC_7424
- ☒ Gloeobacter_violaceus_PCC_7421
- ☒ Microcystis_aeruginosa_NIES_843
- ☒ Nostoc_punctiforme_PCC_73102
- ☒ Prochlorococcus_marinus_MIT_9303
- ☒ Synechococcus_CC_9311
- ☒ Synechococcus_elongatus_PCC_6301
- ☒ Synechococcus_JA_2_3B_a_2_13
- ☒ Synechococcus_PCC_7002
- ☒ Synechocystis_PCC_6803
- ☒ Thermosynechococcus_elongatus_BP1
- ☒ Trichodesmium_erythraeum_IMS_101

☒ Check All

[Retrieve Orthologs](#)

Figure 25: Snapshot showing the list of cyanobacteria and checkboxes for selection of organisms to retrieve bidirectional best hits. The user can select any organism by placing tick mark in the check box of the corresponding cyanobacteria to view bidirectional best hits among the query gene and selected cyanobacteria. The user can also select all the cyanobacteria by placing tick mark in the checkbox 'Check All'.

Gene : **slr2075**

Orthologs

Genome	Bidirectional best hits
Acaryochloris_marina_MBIC_11017	AM1_4412
Anabaena_variabilis_ATCC_29413	Ava_3627
Cyanothece_PCC_7424	PCC7424_1789
Gloeobacter_violaceus_PCC_7421	gvip396
Microcystis_aeruginosa_NIES_843	MAE_46070
Nostoc_punctiforme_PCC_73102	Npun_R0830
Prochlorococcus_marinus_MIT_9303	P9303_05031
Synechococcus_CC_9311	sync_2283
Synechococcus_elongatus_PCC_6301	syc1788_d
Synechococcus_JA_2_3B_a_2_13	CYB_1619
Synechococcus_PCC_7002	SYNPCC7002_A2457
Synechocystis_PCC_6803	slr2075
Thermosynechococcus_elongatus_BP1	tl10186
Trichodesmium_erythraeum_IMS_101	Tery_4326

[Find Cis-acting elements](#)

Click [here](#) to download file.

Figure 26: Snapshot showing the retrieved orthologs for the gene *slr2075* of *Synechocystis*. Upon a single click on the download link orthologs of selected gene(s) can be downloaded in the FASTA format.

Find *Cis*-acting elements

GIBBS MOTIF SAMPLER

Width of the MOTIF :

Expected number of elements for each type :

Near optimal cutoff (0 to 1) :
e.g. 0.5 means 50%

Give seed for random number generator :

MEME

Statistical parameter :

Minimum number of sites for each motif :

Maximum number of sites for each motif:
Maximum is 12

Stop if E-motif value is greater than:

Maximum number of motifs to find:

Width of the motif :

Gap penalty (default it is 11) :

Gap extension penalty (default is 1):

MDSCAN

Width of the motif :

number of top sequences to look for candidate motifs:

Expected bases per motif site (default is none) :

Number of candidate motifs to scan and refine:

Number of top motifs to report at the end:

BIOPROSPECTOR TOOL

Width of the motif :

No. of top sequences to look :

|

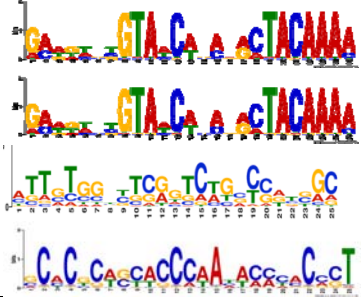
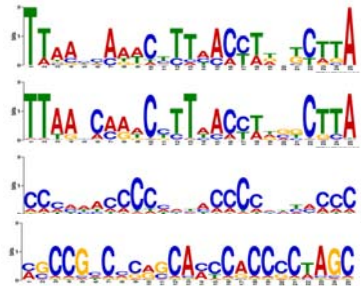
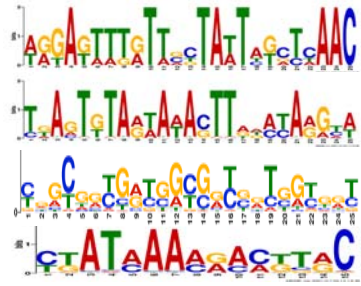
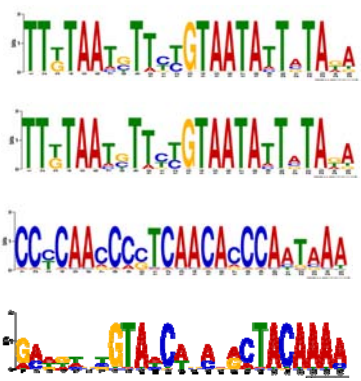
Figure 27: Snapshot showing the parameter file *CyanoCis*. User can change or use the default parameter values for finding *cis*- regulatory elements in the upstreams of selected cyanobacterial genes using four motif prediction tools. User can change any of the parameters, such as ‘width of the motif’, ‘number of motifs to predict’, ‘gap and gap extension penalty’, ‘ZOOPS statistical parameter’, as per the requirement to get better motif predictions by all four integrated tools.

A single click on submit button, generates *cis*-regulatory elements for all the genes which were selected by the user, by four different tools, MEME, Bioprospector, Gibbs motif sampler and MD scan (**Figure 28**).

4.5 Performance analysis of CyanoCis

In order to test the predictions CyanoCis were accurate or not, we selected *slr1756*, *sll0622*, *sll1916*, *smr0009*, *sll0679*, and *sll0822* genes of *Synechocystis* sp. PCC 6803 and identified the *cis*-regulatory elements in their upstreams using CyanoCis. *Cis*-regulatory elements generated by all four integrated motif prediction tools, MEME, Gibbs Motif Sampler, MDScan and Bioproscpector are represented as consensus sequence. The predicted *cis*-acting elements were consistent with the previously identified and experimentally validated ones. The published papers discussing about these motifs are given in column 3. **Table 2** shows the predicted *cis*-regulatory elements which were identified in the clustered-upstreams of the above selected genes by the CyanoCis web-server. In *Synechocystis*, *slr1756* (*glnA*) is regulated by NtcA transcription factor. NtcA acts as both an activator and repressor involved in regulation of a number of nitrogen assimilation genes including *glnA* (Lopatovskaia et al., 2011). The binding site of NtcA is reported to be a tri-nucleotide inverted repeat GTA N(8) TAC. When *glnA* was selected from the 'list of genes' of *Synechocystis*, the same inverted repeat, which was previously identified to be binding site for NtcA was predicted by MEME and Bioproscpector (**Table 2**). In *E. coli* genes involved in the assimilation of phosphorous are regulated by a two-component system consisting of sensory kinase, PhoR and a cognate response regulator, PhoB (Wanner, 1996). SphR of *Synechocystis* is an ortholog of PhoB and its binding site was reported to be TTTAACCA N3 CTTTACTA (Su et al., 2007). The open reading frame, *sll0679*, which codes for a "periplasmic phosphate-binding protein of ABC transporter" was selected for motif prediction tools, the same motif TTTAACAA N3 CTTTACTA was identified by MEME and Bioproscpector (**Table 2**).

Table 2: The *cis*-regulatory elements identified by CyanoCis in the upstreams of selected genes of *Synechocystis* sp. PCC 6803.

Gene	Cis acting elements identified		Reference
	Motif prediction tool	Cis-acting element	
<i>Slr1756 (glnA)</i>	MEME Bioprosector MDScan Gibbs Motif sampler		(Lopatovskaia, et al., 2011)
<i>sll0679 (sphX)</i>	MEME Bioprosector MDScan Gibbs Motif sampler		(Wanner, 1996)
<i>Sll0822 (Arb2)</i>	MEME Bioprosector MDScan Gibbs Motif sampler		(Dutheil, et al., 2012)
<i>Smr0009 (psbN)</i>	MEME Bioprosector MDScan Gibbs Motif sampler		(Seino, et al., 2009)

transcription factor as, when we selected these genes for motif prediction, our web-server predicted a common motif in their upstreams by MEME and Bioprosector (**Table 2**). Gibbs Motif sampler generated a different motif, which is also commonly present in the upstreams of *sll0622* and *sll1916*. These motifs and corresponding transcription factor were not reported previously. As several of the *cis*-regulatory elements predicted by our web-server were exactly matching with the previously experimentally identified ones, we suggest that CyanoCis is suitable for predicting the *cis*-regulatory elements of cyanobacterial genes, in order to facilitate downstream experimental validation of these testable hypotheses.

4.6 Conclusion

CyanoCis is a web server, which can generate orthologs of selected cyanobacterial genes, retrieve upstream DNA sequences of cyCoTs, and identify *cis*-regulatory elements in the upstream of any conserved cyanobacterial gene(s) of interest. The orthologs generated by CyanoCis, upstream DNA sequences of these orthologs, and the *cis*-regulatory elements can be downloaded for further analysis. As CyanoCis is integrated with four different motif prediction tools, the motifs that are commonly identified across multiple methods can be considered as more appropriate targets for further research with confidence. CyanoCis can be accessed from http://jssplab.uohyd.ac.in/Cyanocis_2.

Chapter 5

Objective 4

UpCoT: an integrative pipeline tool for clustering upstream DNA sequences of orthologous genes in prokaryotic genomes.

UpCoT: an integrative pipeline tool for clustering upstream DNA sequences of orthologous genes in prokaryotic genomes

Summary

In this chapter we describe an integrated pipeline tool UpCoT. The first step of the phylogenetic foot printing (i.e., grouping upstreams of orthologous genes in closely related organisms) requires use of several computational tools and processes. UpCoT is an integrative pipeline tool developed by automating the series of steps involved in prediction of *cis*-regulatory elements for ease and quick extraction of upstreams of homologous genes. UpCoT generates orthologs by bidirectional best hit method, cluster of ortholog groups (tgCoGs), cluster of transcription units (tgCoTs), and clustered-upstream DNA sequences of any target prokaryotic genome. The inputs of UpCoT are *.faa, *.fna, *.ptt files of a target and reference genomes which are provided along with the UpCoT package. A total of 1840 prokaryotic genomes are available in the web page of UpCoT for selection to generate clustered-upstream DNA sequences. The output generated by UpCoT contains tgCoG file and the clustered-upstream DNA sequences of each CoT of the target genome in FASTA format. These clustered-upstream DNA sequences can be used by any motif prediction tool, such as MEME, Bio-prospector, Gibbs motif sampler, MDscan for prediction of *cis*-regulatory elements. We tested the performance of UpCoT by selecting the genome of *Synechocystis* sp PCC 6803 as the target and 13 different cyanobacterial genomes as reference. UpCoT generated 2578 clustered upstream DNA sequences. Out of these clustered-upstream DNA sequences, the output of *groES*, *ycf24* and *nirA* were used for *cis*-regulatory element prediction. The results were consistent with the experimentally identified *cis*-regulatory elements. Therefore, UpCoT is a reliable and accurate

automated software for prediction of orthologs, clusters of transcriptional units, and clustered-upstream DNA sequences of a selected prokaryotic genome. UpCoT can be downloaded from [*http://jssplab.uohyd.ac.in/upcot/*](http://jssplab.uohyd.ac.in/upcot/).

5. Introduction

With the advent of fast and next generation automated DNA sequencing technologies, a number of microbial genomes have been sequenced during the past decade and the sequence information is available in various genome databases. Identification of *cis*-regulatory elements and the trans-acting factors of a sequenced genome is one of the major challenges to computational biologists for building a global gene regulatory network. Phylogenetic footprinting is one of the widely accepted computational method for predicting *cis*-regulatory elements for a given genome in question (Hardison, 2000). This method can be considered as a two step process. The first step involves, identification of orthologs in the reference genomes, for each protein of a target genome by bidirectional best hit method, prediction of transcriptional units of target and reference genomes, generation of cluster of transcriptional units (CoTs), and finally clustering of upstream DNA sequences based on the generated CoT data (Wels et al., 2006). The second step involves scanning for conserved DNA elements in the clustered-upstream DNA sequences of a given CoT. Various computational tools, such as MEME, Bioprosector, Gibbs sampler, MDScan are used for predicting conserved DNA elements in a given set of DNA sequences and can be represented in the form of consensus pattern (Bailey and Elkan, 1994; Neuwald et al., 1995; Liu et al., 2001; Liu et al., 2002; Mrazek, 2009). There are many computational tools to perform the second step of phylogenetic foot printing but, they are not available to perform the first step. Further, the first step by itself is a multi-step process and requires lengthy computational procedure. On the other hand, the number of microbial genomes being sequenced is constantly increasing and demands for the development of an automated tool. Developing such a tool would facilitate the biologists to work easily on any microbial genome for quick generation of clustered-upstream DNA sequences for the target genome in question.

Keeping the above facts in view, we developed an automated integrated pipeline called, UpCoT, which generates the orthologs for proteins of target organism (tgCoGs), generates clusters of transcription units (tgCoTs), and cluster the upstream DNA sequences of tgCoTs. The output of the UpCoT can be directly used for prediction of *cis*-regulatory elements using any computational tool of user's choice.

5.1 Materials and Methods

5.1.1 Design of UpCoT interface

UpCoT web interface was designed using HTML, PHP and javascript to select and retrieve the genomes for analysis by UpCoT package. The *.faa, *.fna, and *.ptt files of 1840 prokaryotic genomes were downloaded from NCBI (<ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria>) and incorporated in the web server, where UpCoT package has been maintained. User can select any of these genomes as target and reference files as input to UpCoT package. UpCoT package along with selected genomes for analysis can be downloaded from the web link, <http://jssplab.uohyd.ac.in/upcot>. The UpCoT package was developed using Perl programming language and is compatible for Windows and Linux operating systems.

5.2 Description and accessibility of UpCoT web interface

The web interface contains 'Home', 'Help' and 'Contact' links below the header. A brief introduction about the UpCoT is given in the homepage (**Figure 29**). A single click on 'Select Genomes' link navigates to a new web page displaying the list of prokaryotic genomes for selection of target genome (**Figure 30**).

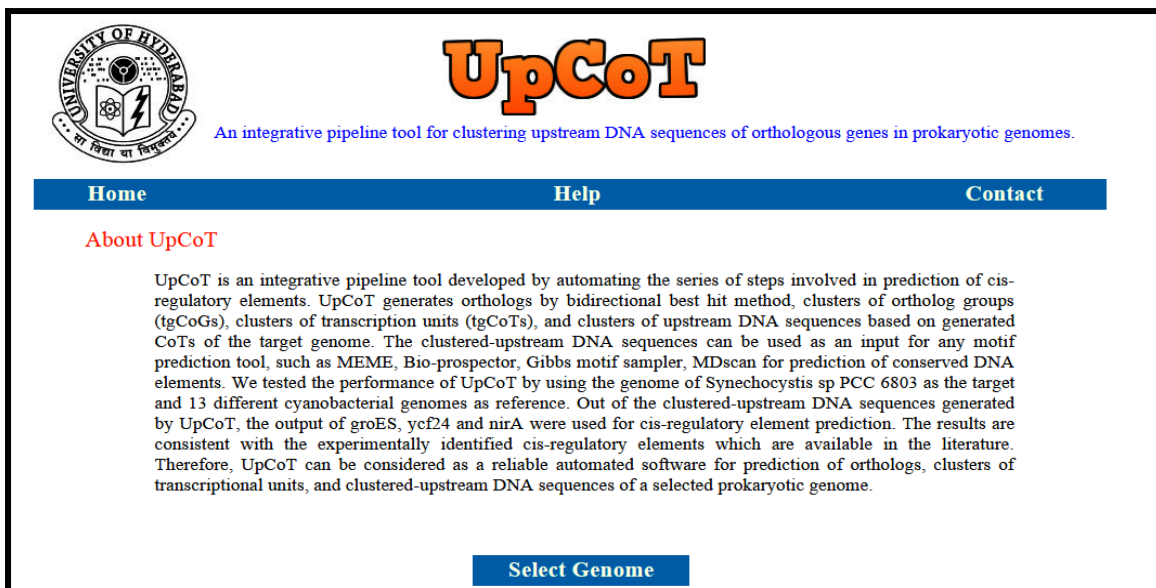
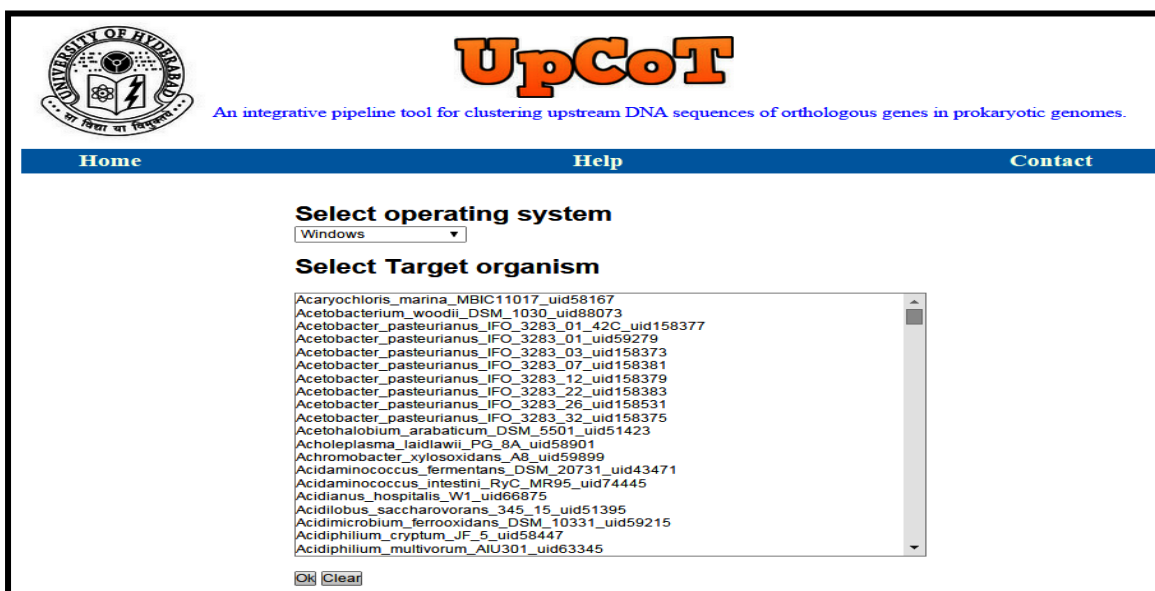


Figure 29: Snapshot showing the home page of UpCoT. The homepage links for 'Select Genome', 'Home', 'Help' and 'Contact'. Upon clicking on the link 'Select Genome', a new page appears where the user can select the target genome on interest.

User needs to select the operating system (either windows or linux operating system) from the dropdown box provided, for downloading suitable executables to run the UpCoT package. Upon selecting the operating system and the target organism, a new page appears and prompts the user for selecting reference genomes (**Figure 31**). Any number of reference genomes can be selected for analysis. After the selection of both target and reference genomes, user can download the UpCoT package by a single click on 'Download UpCoT' tab (**Figure 32**). The UpCoT windows version needs a supporting software package GNU on windows to be installed, which is provided along with UpCoT package. UpCoT package also contains 'settings.txt', 'README.txt', 'target_genome.txt', and 'reference_genomes.txt' files (**Figure 33**). The "bin" directory provided in the UpCoT package contains Perl programs developed for performing different tasks such as blastP, extraction of top scoring hits, prediction of bidirectional best hits, counting the number of orthologs, upstream sequence retrieval, and generating clustered-upstream sequences. User may change the parameters, such as E-value (default, $d = 1e^{-3}$), orthologs count (default = 4),

computer configuration (default = 64 bit), installation path of GNU on window (in case of windows user), minimum upstream length (min_UP length default = 50) and maximum upstream length (max_UP length default = 350 bp), in "settings.txt" file provided in the package (Figure 34).



The screenshot shows the UpCoT web interface. At the top left is the University of Hyderabad logo. The title 'UpCoT' is in large orange letters, followed by the subtitle 'An integrative pipeline tool for clustering upstream DNA sequences of orthologous genes in prokaryotic genomes.' Below this is a navigation bar with 'Home', 'Help', and 'Contact' links. The main content area has two sections: 'Select operating system' with a dropdown menu set to 'Windows', and 'Select Target organism' with a scrollable list of 20 prokaryotic organisms. At the bottom of the list are 'OK' and 'Clear' buttons.

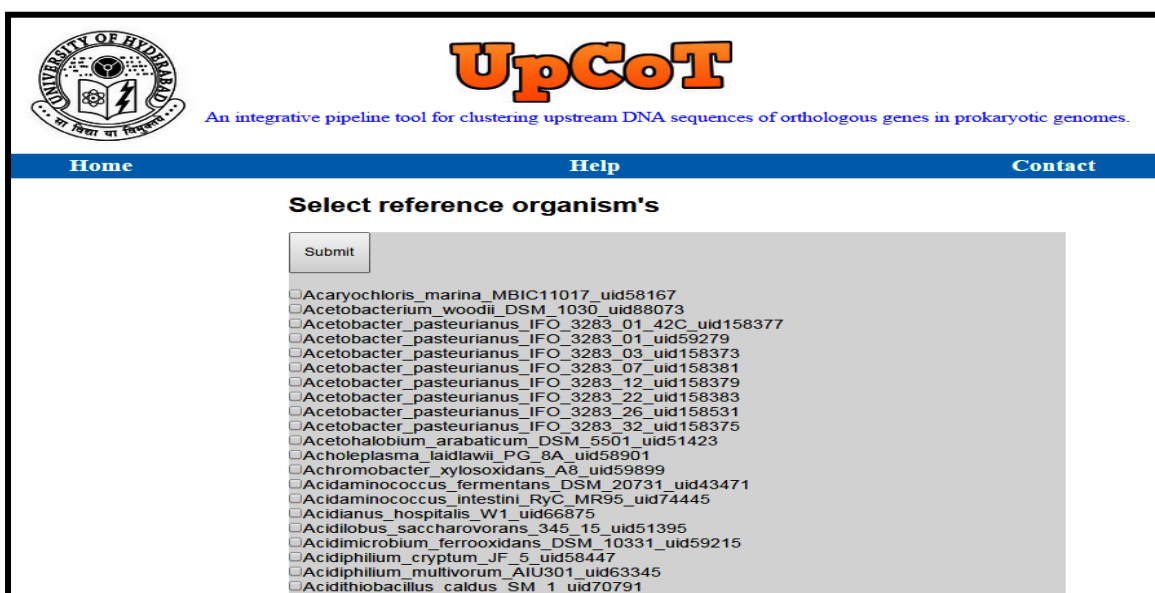
Select operating system
Windows

Select Target organism

- Acaryochloris_marina_MBIC11017_uid58167
- Acetobacterium_woodii_DSM_1030_uid88073
- Acetobacter_pasteurianus_IFO_3283_01_42C_uid158377
- Acetobacter_pasteurianus_IFO_3283_01_uid59279
- Acetobacter_pasteurianus_IFO_3283_03_uid158373
- Acetobacter_pasteurianus_IFO_3283_07_uid158381
- Acetobacter_pasteurianus_IFO_3283_12_uid158379
- Acetobacter_pasteurianus_IFO_3283_22_uid158383
- Acetobacter_pasteurianus_IFO_3283_26_uid158531
- Acetobacter_pasteurianus_IFO_3283_32_uid158375
- Acetohalobium_arabaticum_DSM_5501_uid51423
- Acholeplasma_laidlawii_PG_8A_uid58901
- Achromobacter_xylosoxidans_A8_uid59899
- Acidaminococcus_fermentans_DSM_20731_uid43471
- Acidaminococcus_intestini_RyC_MR95_uid74445
- Acidianus_hospitalis_W1_uid66875
- Acidilobus_saccharovorans_345_15_uid51395
- Acidimicrobium_ferrooxidans_DSM_10331_uid59215
- Acidiphilium_cryptum_JF_5_uid58447
- Acidiphilium_multivorum_AIU301_uid63345

OK Clear

Figure 30: Snapshot showing the options to select the type of operating system and list of target organisms.



The screenshot shows the UpCoT web interface. At the top left is the University of Hyderabad logo. The title 'UpCoT' is in large orange letters, followed by the subtitle 'An integrative pipeline tool for clustering upstream DNA sequences of orthologous genes in prokaryotic genomes.' Below this is a navigation bar with 'Home', 'Help', and 'Contact' links. The main content area has a section titled 'Select reference organism's' with a 'Submit' button and a scrollable list of 20 prokaryotic organisms, each preceded by a checkbox.

Select reference organism's

Submit

- ☐ Acaryochloris_marina_MBIC11017_uid58167
- ☐ Acetobacterium_woodii_DSM_1030_uid88073
- ☐ Acetobacter_pasteurianus_IFO_3283_01_42C_uid158377
- ☐ Acetobacter_pasteurianus_IFO_3283_01_uid59279
- ☐ Acetobacter_pasteurianus_IFO_3283_03_uid158373
- ☐ Acetobacter_pasteurianus_IFO_3283_07_uid158381
- ☐ Acetobacter_pasteurianus_IFO_3283_12_uid158379
- ☐ Acetobacter_pasteurianus_IFO_3283_22_uid158383
- ☐ Acetobacter_pasteurianus_IFO_3283_26_uid158531
- ☐ Acetobacter_pasteurianus_IFO_3283_32_uid158375
- ☐ Acetohalobium_arabaticum_DSM_5501_uid51423
- ☐ Acholeplasma_laidlawii_PG_8A_uid58901
- ☐ Achromobacter_xylosoxidans_A8_uid59899
- ☐ Acidaminococcus_fermentans_DSM_20731_uid43471
- ☐ Acidaminococcus_intestini_RyC_MR95_uid74445
- ☐ Acidianus_hospitalis_W1_uid66875
- ☐ Acidilobus_saccharovorans_345_15_uid51395
- ☐ Acidimicrobium_ferrooxidans_DSM_10331_uid59215
- ☐ Acidiphilium_cryptum_JF_5_uid58447
- ☐ Acidiphilium_multivorum_AIU301_uid63345
- ☐ Acidithiobacillus_caldus_SM_1_uid70791

Figure 31: Snapshot of the page showing the list of reference organisms the user wishes to select. The user can select any number of the organisms and then click on “submit” to proceed further.

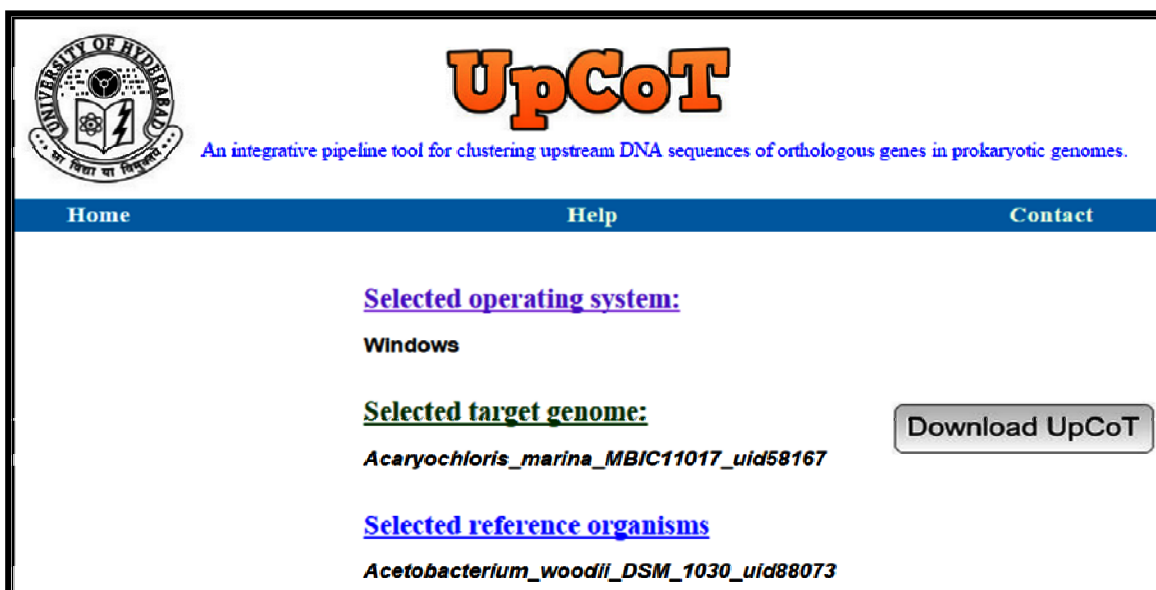


Figure 32: Snapshot of the page showing information at glance. The user can obtain the information of selected operating system, selected target organism and the list of reference organisms. Upon clicking on the “Download UpCoT” the selected suitable UpCoT package with all Perl programs and with the target and reference organisms *.faa, *.ptt, *.fna are downloaded in the form of a zip file.

5.3 UpCoT Output

UpCoT identifies the orthologs for each protein of target genome by bidirectional best hit method (BDBH) in the given reference genomes using BlastP (Altschul et al., 1990). After performing BDBH, UpCoT generates clusters of orthologs groups for target genome (tgCoGs) based on the orthologs count. For example, if the orthologs count is set to 4, tgCoGs containing four orthologs or above will be selected for further analysis.

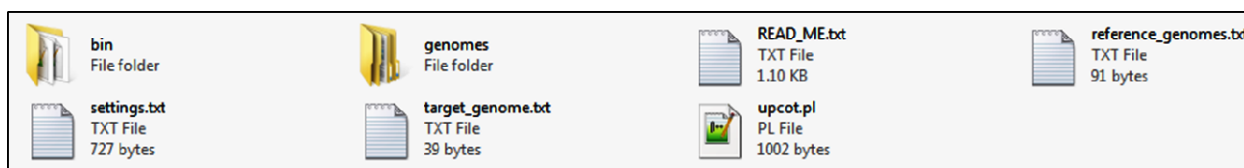
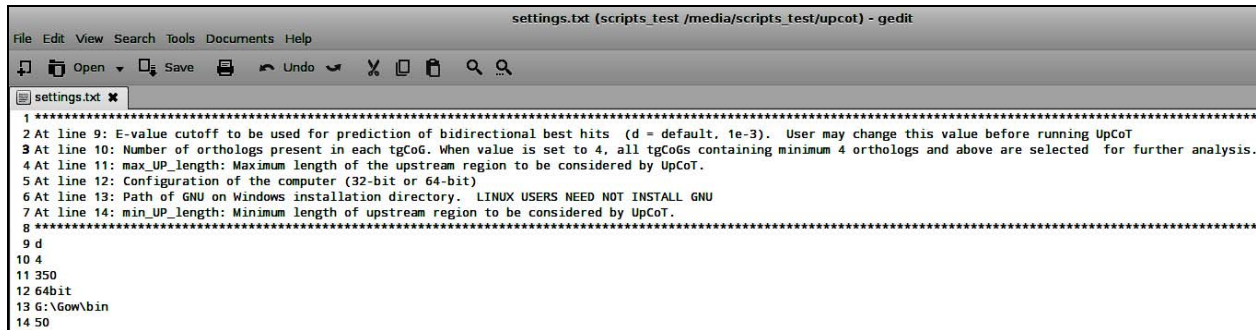


Figure 33: Snapshot showing the directories, and files present in the UpCoT package. The “bin” directory contains Perl programs needed for running of UpCoT. The ‘genomes’ directory contains *.faa, *.fna, *.ptt files of selected target and reference genomes. ‘Read_Me.txt’ provides the instructions about how to use UpCoT package. The file ‘settings.txt’ provides the input parameters. The ‘upcot.pl’ is the main file which invokes all the Perl programs that are present in “bin” directory.



```
1 *****
2 At line 9: E-value cutoff to be used for prediction of bidirectional best hits (d = default, 1e-3). User may change this value before running UpCoT
3 At line 10: Number of orthologs present in each tgCoG. When value is set to 4, all tgCoGs containing minimum 4 orthologs and above are selected for further analysis.
4 At line 11: max_UP_length: Maximum length of the upstream region to be considered by UpCoT.
5 At line 12: Configuration of the computer (32-bit or 64-bit)
6 At line 13: Path of GNU on Windows installation directory. LINUX USERS NEED NOT INSTALL GNU
7 At line 14: min_UP_length: Minimum length of upstream region to be considered by UpCoT.
8 *****
9 d
10 4
11 350
12 64bit
13 G:\Gow\bin
14 50
```

Figure 34: Snapshot showing the file contents of 'settings.txt'. E-value cutoff to be used for prediction of bidirectional best hits ($d = \text{default}, 1e^{-3}$). the values of the settings file may be changed by the user. Number of orthologs present in each tgCoG, by default the value is set to 4. When value is set to 4, all tgCoGs containing minimum 4 orthologs and above are selected for further analysis. Max_UP_length, the maximum length of the upstream region to be considered by UpCoT. By default Max_UP_length is set to 350bp. Configuration of the computer (32-bit or 64-bit). Path of GNU on Windows installation directory. Min_UP_length; Minimum length of upstream region to be considered by UpCoT. By default its value is set to 50bp.

A directory named as "tgCoG_protein_sequences" is generated containing FASTA files of tgCoG protein sequences. In addition, UpCoT generates clusters of transcriptional units (tgCoTs) by reading the tgCoG files and also based on the length of their corresponding upstream DNA sequences. The minimum length of the upstream region (min_UP length default = 50) and the maximum length of the upstream region (max_UP length default = 350). It excludes the upstreams of the open reading frames, which are less than the defined nucleotide length. Reports suggest that the genes possessing an upstream region less than 40 to 50 bp are to be excluded from the computational prediction of *cis*-regulatory elements, we have set the minimum default integer value as 50 bp in the "settings.txt" file (**Figure 34**) (Salgado et al., 2000; Conlan et al., 2005; Liu et al., 2008). User may change the minimum length as per the requirement. When the actual length of the upstream region is greater than the minimum default length or user-defined minimum length, the program selects 350 bp upstream region of an ORF. When the upstream intergenic region is longer than 350 bp, the UpCoT considers only 350 bp upstream region, as the max_UP length is set to 350 bp. User may also change the max_length

according to the requirement. Subsequently, UpCoT extracts and clusters the upstream DNA sequences of tgCoTs based on default or user defined integer values as upstream length given in "settings.txt" file. Upon completion of the whole process, a directory named "tu_upstreams" appears in the working directory of the UpCoT. This directory contain multiple text files, each with clustered-upstream DNA sequences of a tgCoT. Each text file is named with ORF number of the target gene. User can submit these upstream sequences for any motif prediction tool for identifying *cis*-regulatory elements. The entire work flow of UpCoT including inputs, the processes, and the outputs are depicted in (Figure 35).

5.4 Methodology used for testing UpCoT

We used the genome of *Synechocystis* sp. PCC 6803 (hereafter *Synechocystis*) as the target and *Acaryochloris marina* MBIC 11017, *Synechococcus* CC 9311, *Anabaena variabilis* ATCC 29413, *Synechococcus elongatus* PCC 6301, *Cyanothece* PCC 7424, *Synechococcus* JA 2 3B a 2 13, *Gloeobacter violaceus* PCC 7421, *Synechococcus* PCC 7002, *Microcystis aeruginosa* NIES 843, *Nostoc punctiforme* PCC 73102, *Thermosynechococcus elongatus* BP1, *Prochlorococcus marinus* MIT 9303, *Trichodesmium erythraeum* IMS 101 as reference genomes. We used the default E-value ($d = 1e^{-3}$), ortholog count as 4 and min_UP length as 50 and max_UP length as 500 bp for testing UpCoT. From the output generated by UpCoT, the text files with names *Slr2075*, *Slr0074* and *Slr0898* were selected from 'tu_upstreams' directory and submitted for the prediction of *cis*-regulatory elements using standalone versions of MEME, Gibbs Motif Sampler, MDScan and Bioprosector.

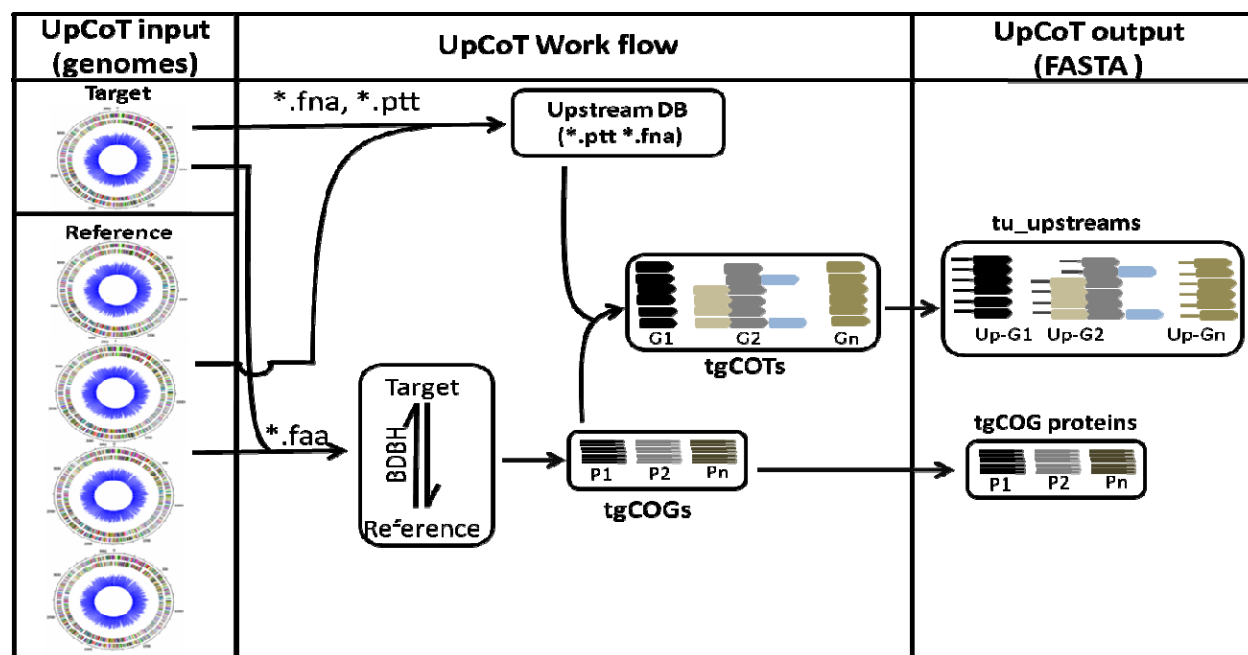


Figure 35: The schematic representation of the UpCoT input, UpCoT work flow and UpCoT output. The inputs for UpCoT are *.faa, *.fna, *.ptt files of target and reference genomes of user's choice. UpCoT uses these files to generate tgCoGs by Bidirectional best hit method (BDBH) and the clusters of transcriptional units (tgCoTs). UpCoT groups the upstreams of each gene of a tgCoT to generate clustered-DNA upstreams of that tgCoT. All clustered-DNA upstreams of each tgCoT are saved in to 'tu-upstreams' directory. Each output file is a text file named with 'Up-ORF id' of the target organism. UpCoT also generates the tgCoG protein sequences as text files. G1, tgCoT of gene 1; P1, tgCoG of gene 1; Up-Gn, clustered-upstream sequences of gene 'n' of a target genome.

5.5 Results and discussion

5.5.1 Performance analysis of UpCoT

The target genome, *Synechocystis* has 3172 open reading frames that code for proteins involved in various cellular processes and unknown proteins. Out of 3172 proteins, UpCoT has generated 2578 tgCoGs, each containing minimum four orthologs. This shows that 81 % of *Synechocystis* proteins are present in at least four selected cyanobacterial species. Orthologs identified by UpCoT for the selected proteins were retrieved from 'tgCoG_protein_sequences' directory and tested for accuracy. **Table 3** shows the selected proteins and their orthologs along with their functional annotation. From **Table 3**, it is clear that the orthologs identified by UpCoT for the proteins Slr2075 (GroES), Slr0074 (Ycf24), Slr0898

(NirA), Ssl2598 (PsbH), Smr0009 (PsbN), Sll0851 (PsbC) and Sll0894 (PsbD) are accurate because their annotations are same as given in NCBI genome database.

5.5.2 Analysis of clustered-upstream DNA sequences for selected tgCoTs

UpCoT has generated 2578 text files each containing clustered-upstream DNA sequences of a tgCoT. Out of which, the clustered upstream DNA sequences of *slr2075*, *slr0074* and *slr0898* were submitted for *cis*-regulatory element prediction, as the regulatory elements for these genes were previously experimentally demonstrated to be the target sites for known transcription factors. The clustered-upstreams were submitted to four different motif prediction tools as described in the materials and methods. (**Table 4**) shows the predicted *cis*-regulatory elements which were identified in the clustered-upstreams of the above selected tgCoTs. In *Synechocystis* the gene *slr2075* encodes for co-chaperonin GroES. HrcA, a transcriptional repressor has been reported regulate the expression of the *groESL* operon by binding to a 9-bp inverted repeat TTAGCACTC [N9] GAGTGCTAA (Zuber and Schumann, 1994; Nakamoto et al., 2003). When the clustered-upstream sequences of *slr2075*-tgCoT was submitted as input to motif prediction tools the same inverted repeat was predicted by MEME, Gibbs motif sampler and Bioprosppector (**Table 4**). The *SufR* is a negative transcriptional regulator of *sufBCDS* operon in *Synechocystis*. *SufR* binds to *cis*-regulatory element, CAAC-N6-GTTG located between the divergently transcribed *sufR* gene and the *sufBCDS* operon, and acts as a repressor of the *sufBCDS* operon and as an auto-regulator of its own gene, *sufR* (Wang et al., 2004). Motif prediction tools MEME, MDScan and Bioprosppector generated the same element upon submission of clustered-upstream sequences of *slr0074*-tgCoT (**Table 4**). A number of nitrogen assimilation genes are regulated by the global transcriptional regulator NtcA, that acts as both an activator and repressor (Aichi et al., 2001). The binding site of NtcA is reported to be a tri-

nucleotide inverted repeat GTA N(8) TAC. The ORF, *slr0898* codes for Ferredoxin-nitrite reductase (NirA) in *Synechocystis*. MEME and Bioprosector tool has predicted the NtcA binding site in the clustered-upstream DNA sequences of tgCoT-*slr0898* (**Table 4**). Thus, based on the identification of experimentally validated *cis*-regulatory elements for clustered upstreams of tgCoTs, we suggest that UpCoT is suitable for extracting and clustering of upstreams for any group of microbial genomes with accuracy and can be used for phylogenetic foot printing, promoter prediction, sRNA mapping and TSS prediction.

4.6 Conclusion

UpCoT is an automated software that can perform prediction of bidirectional best hits, clusters of transcriptional units (tgCoTs) and grouping of upstream DNA sequences for the predicted tgCoTs in a single step. It can be used as a tool by biologists to work on available microbial genomes for prediction of *cis*-regulatory elements using phylogenetic foot printing. UpCoT can be downloaded from <http://jssplab.uohyd.ac.in/upcot/>.

Table 3: Orthologs identified by UpCoT for selected proteins of target organism, *Synechocystis* sp. PCC6803.

	Orthologous proteins identified for selected proteins of target organism, <i>Synechocystis</i> sp. PCC6803 by UpCoT						
<i>Synechocystis</i> sp. PCC 6803	Slr2075 (GroES) Co-chaperonin	Slr0074 (Ycf24) Cysteine desulfurase activator complex subunit	Slr0898 (NirA) Ferredoxin-nitrite reductase	Ssl2598 (PsbH) Photosystem II reaction center protein H	Smr0009 (PsbN) Photosystem II reaction center protein N	Sll0851 (PsbC) Photosystem II CP43 protein	Sll0849 (PsbD) Photosystem II D2 protein
<i>Acaryochloris marina</i> MBIC 11017	Am1_4412	Am1_1224	Am1_2984	Am1_1677	Am1_5511	Am1_1084	Am1_4084
<i>Anabaena variabilis</i> ATCC 29413	Ava_3627	Ava_0424	Ava_4539	Ava_2220	Ava_4451	Ava_1243	Ava_2512
<i>Cyanothece</i> PCC 7424	Pcc7424_1789	Pcc7424_4729	Pcc7424_1683	Pcc7424_1517	Pcc7424_4233	Pcc7424_0578	Pcc7424_2974
<i>Gloeobacter violaceus</i> PCC 7421	Gvip396	Gvip196	Gvip212	Gsl1716	Gvip411	Gvip319	Gvip318
<i>Microcystis aeruginosa</i> NIES 843	Mae_46070	Mae_23090	Mae_18410	Mae_11070	Mae_36550	Mae_41150	Mae_41160
<i>Nostoc punctiforme</i> PCC 73102	Npun_r0830	Npun_f4822	Npun_r1528	Npun_f1088	Npun_r4314	Npun_r3636	Npun_f4553
<i>Prochlorococcus marinus</i> MIT 9303	P9303_05031	P9303_03021	P9303_29861	P9303_18181	P9303_24631	P9303_08421	P9303_08431
<i>Synechococcus</i> CC 9311	Sync_2283	Sync_2483	Sync_2898	Sync_1909	Syc_0309	Sync_0896	Sync_2586
<i>Synechococcus elongatus</i> PCC 6301	Syc1788_d	Syc2356_c	Syc0310_d	Syc0977_c	Syc1289_d	Syc0872_c	Syc0873_c
<i>Synechococcus</i> JA 2 3B a 2 13	Cyb_1619	Cyb_1405	Cyb_0034	Not identified	Cyb_1372	Cyb_0853	Cyb_1736
<i>Synechococcus</i> PCC 7002	Synpcc7002_a2457	Synpcc7002_a1814	Synpcc7002_a1827	Not identified	Synpcc7002_a0809	Synpcc7002_a1559	Synpcc7002_a2199
<i>Thermosynechococcus elongatus</i> BP1	Tll0186	Tll0490	Tlr1349	Tsr0149	Tsr1387	Tlr1631	Tlr1630
<i>Trichodesmium erythraeum</i> IMS 101	Tery_4326	Tery_4355	Tery_1068	not identified	Tery_2867	Tery_0513	Tery_1230

Synechocystis was used as target and other selected cyanobacterial species were used as reference organisms. Functional annotation is given below the name and is based on NCBI genome database (<ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria>).

Table 4: *cis*-regulatory elements identified in the clustered-upstreams of selected tgCoTs generated by UpCoT

Clustered-upstreams of tgCoT used for motif prediction	<i>Cis</i> -regulatory element predicted in the clustered-upstreams of tgCoT		Reference in which motif was experimentally proven to be <i>cis</i> -regulatory element
	Motif prediction tool	Predicted <i>cis</i> -regulatory element	
<i>Up_slr2075_CoT</i>	MEME Gibbs Motif Sampler MDScan Bioprospesor	 NOT IDENTIFIED 	(Nakamoto, et al., 2003)
<i>Up_slr0074_CoT</i>	MEME Gibbs Motif Sampler MDScan Bioprospesor	 	(Wang, et al., 2004)
<i>Up_slr0898_CoT</i>	MEME Gibbs Motif Sampler MDScan Bioprospesor	 	(Aichi, et al., 2001)

The clustered-upstreams of slr2075-tgCoT (*Up_slr2075_CoT*), slr0074-tgCoT (*Up_slr0074_CoT*) and slr0898-tgCoT (*Up_slr0898_CoT*) were submitted to MEME, Gibbs Motif Sampler, MDScan and Bioprospesor tools for identifying *cis*-regulatory elements. The predicted *cis*-regulatory elements are shown as a consensus sequence. The predicted conserved sequences were consistent with the previously published and experimentally validated *cis*-regulatory elements.

References

- Aichi M, Takatani N, Omata T** (2001) Role of NtcB in activation of nitrate assimilation genes in the cyanobacterium *Synechocystis* sp. strain PCC 6803. *J Bacteriol* **183**: 5840-5847
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ** (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403-410
- Arun PPS, Subhashini M, Santhosh C, Krishna PS, Prakash JS** (2013) Prediction of Cis Regulatory Elements in the Genome of *Synechococcus Elongatus* PCC 6301. *In* Photosynthesis Research for Food, Fuel and the Future. Springer, pp 369-373
- 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
- Ashokan KV, Pillai MM** (2008) In silico characterization of silk fibroin protein using computational tools and servers. *Asian Journal of experimental science* **22**: 265-274
- Axmann IM, Kensche P, Vogel J, Kohl S, Herzel H, Hess WR** (2005) Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biol* **6**: R73
- Bailey TL, Elkan C** (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28-36
- Bhattacharjee A, Choudhury H, Maheswari U, Joshi SR** (2008) Insilico prediction of structural and functional aspects of a hypothetical protein of *Arabidopsis thaliana* (L) Heynh. *Advanced Biotech* **7**: 14-16
- Breitaudeau A, Coste F, Humily F, Garczarek L, Le Corguille G, Six C, Ratin M, Collin O, Schluchter WM, Partensky F** (2013) CyanoLyase: a database of phycobilin lyase sequences, motifs and functions. *Nucleic Acids Res* **41**: D396-401
- Bureau TE, Wessler SR** (1994) Mobile inverted-repeat elements of the Tourist family are associated with the genes of many cereal grasses. *Proc Natl Acad Sci U S A* **91**: 1411-1415
- Casacuberta JM, Santiago N** (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene* **311**: 1-11
- Chen Y, Zhou F, Li G, Xu Y** (2009) MUST: a system for identification of miniature inverted-repeat transposable elements and applications to *Anabaena variabilis* and *Haloquadratum walsbyi*. *Gene* **436**: 1-7
- Chou PY, Fasman GD** (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol* **47**: 45-148
- Conlan S, Lawrence C, McCue LA** (2005) *Rhodospseudomonas palustris* regulons detected by cross-species analysis of alphaproteobacterial genomes. *Appl Environ Microbiol* **71**: 7442-7452
- Creighton TE** (1993) Proteins: structures and molecular properties. Macmillan
- Delwiche C, Palmer J** (1997) The origin of plastids and their spread via secondary symbiosis. *In* D Bhattacharya, ed, *Origins of Algae and their Plastids*, Vol 11. Springer Vienna, pp 53-86
- Dutheil J, Saenkhram P, Sakr S, Leplat C, Ortega-Ramos M, Bottin H, Cournac L, Cassier-Chauvat C, Chauvat F** (2012) The AbrB2 autorepressor, expressed from an atypical promoter, represses the hydrogenase operon to regulate hydrogen production in *Synechocystis* strain PCC6803. *J Bacteriol* **194**: 5423-5433
- Edwards DJ, Holt KE** (2013) Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microb Inform Exp* **3**: 2
- Fink AL** (1998) Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold Des* **3**: R9-23
- Finn RD, Miller BL, Clements J, Bateman A** (2014) iPFam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Res* **42**: D364-373
- Frishman D, Argos P** (1997) Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* **27**: 329-335

- Gan HM, Hudson AO, Rahman AY, Chan KG, Savka MA** (2013) Comparative genomic analysis of six bacteria belonging to the genus *Novosphingobium*: insights into marine adaptation, cell-cell signaling and bioremediation. *BMC Genomics* **14**: 431
- Ganley AR, Kobayashi T** (2007) Phylogenetic footprinting to find functional DNA elements. *Methods Mol Biol* **395**: 367-380
- Gao XY, Zhi XY, Li HW, Klenk HP, Li WJ** (2014) Comparative genomics of the bacterial genus *Streptococcus* illuminates evolutionary implications of species groups. *PLoS One* **9**: e101229
- Gill SC, von Hippel PH** (1989) Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem* **182**: 319-326
- Hardison RC** (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet* **16**: 369-372
- Harrison RG** (2000) Expression of soluble heterologous proteins via fusion with NusA in *Novations* **11**: 4-7
- Hernandez-Prieto MA, Futschik ME** (2012) CyanoEXpress: A web database for exploration and visualisation of the integrated transcriptome of cyanobacterium *Synechocystis* sp. PCC6803. *Bioinformatics* **8**: 634-638
- Hongbin Liu HAN, Lisa Campbell** (1997) *Prochlorococcus* growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean *Aquatic Microbial Ecology* **12**: 39-47
- Idicula-Thomas S, Balaji PV** (2005) Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*. *Protein Sci* **14**: 582-592
- Ikai A** (1980) Thermostability and aliphatic index of globular proteins. *J Biochem* **88**: 1895-1898
- John P** (1998) *Nitrogen Fixation*, Ed 3rd. Cambridge University Press
- Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S** (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions (supplement). *DNA Res* **3**: 185-209
- Kanesaki Y, Yamamoto H, Paithoonrangsarit K, Shoumskaya M, Suzuki I, Hayashi H, Murata N** (2007) Histidine kinases play important roles in the perception and signal transduction of hydrogen peroxide in the cyanobacterium, *Synechocystis* sp. PCC 6803. *Plant J* **49**: 313-324
- Kantardjieff KA, Rupp B** (2004) Protein isoelectric point as a predictor for increased crystallization screening efficiency. *Bioinformatics* **20**: 2162-2168
- Kelly L, Huang KH, Ding H, Chisholm SW** (2012) ProPortal: a resource for integrated systems biology of *Prochlorococcus* and its phage. *Nucleic Acids Res* **40**: D632-640
- Kim WY, Kang S, Kim BC, Oh J, Cho S, Bhak J, Choi JS** (2008) SynechoNET: integrated protein-protein interaction database of a model cyanobacterium *Synechocystis* sp. PCC 6803. *BMC Bioinformatics* **9 Suppl 1**: S20
- Kiraga J, Mackiewicz P, Mackiewicz D, Kowalczyk M, Biecek P, Polak N, Smolarczyk K, Dudek MR, Cebrat S** (2007) The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. *BMC Genomics* **8**: 163
- Kopito RR** (2000) Aggresomes, inclusion bodies and protein aggregation. *Trends Cell Biol* **10**: 524-530
- Koschorreck M, Fischer M, Barth S, Pleiss J** (2005) How to find soluble proteins: a comprehensive analysis of alpha/beta hydrolases for recombinant expression in *E. coli*. *BMC Genomics* **6**: 49

- Kumwenda B, Litthauer D, Bishop OT, Reva O** (2013) Analysis of protein thermostability enhancing factors in industrially important thermus bacteria species. *Evol Bioinform Online* **9**: 327-342
- Kyte J, Doolittle RF** (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**: 105-132
- Levene PA SH** (1923) Calculation of iso-electric points. *Journal of Biological Chemistry* **55**: 801-813
- Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ** (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res* **34**: W32-37
- Liu J, Xu X, Stormo GD** (2008) The cis-regulatory map of Shewanella genomes. *Nucleic Acids Res* **36**: 5376-5390
- Liu X, Brutlag DL, Liu JS** (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*: 127-138
- Liu XS, Brutlag DL, Liu JS** (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol* **20**: 835-839
- Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ** (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol* **10**: 599-606
- Lopatovskaia KV, Seliverstov AV, Liubetskii VA** (2011) [NtcA- and NtcB-regulons in cyanobacteria and Rhodophyta chloroplasts]. *Mol Biol (Mosk)* **45**: 570-574
- Los DA, Suzuki I, Zinchenko VV, Murata N** (2008) Stress responses in Synechocystis: regulated genes and regulatory systems. Caister Academic Press: Norfolk, UK
- Maharaj Y, Soliman ME** (2013) Identification of novel gyrase B inhibitors as potential anti-TB drugs: homology modelling, hybrid virtual screening and molecular dynamics simulations. *Chem Biol Drug Des* **82**: 205-215
- Mao F, Dam P, Chou J, Olman V, Xu Y** (2009) DOOR: a database for prokaryotic operons. *Nucleic Acids Res* **37**: D459-463
- Mason Gamer RJ** (2007) Multiple homoplasious insertions and deletions of a Triticeae (Poaceae) DNA transposon: a phylogenetic perspective. *BMC Evol Biol* **7**: 92
- Mathura VS, Kolippakkam D** (2005) APDbase: Amino acid Physico-chemical properties Database. *Bioinformatics* **1**: 2-4
- McGuffin LJ, Bryson K, Jones DT** (2000) The PSIPRED protein structure prediction server. *Bioinformatics* **16**: 404-405
- Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D** (2009) Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biol Direct* **4**: 13
- Mrazek J** (2009) Finding sequence motifs in prokaryotic genomes--a brief practical guide for a microbiologist. *Brief Bioinform* **10**: 525-536
- Nakamoto H, Suzuki M, Kojima K** (2003) Targeted inactivation of the hrcA repressor gene in cyanobacteria. *FEBS Lett* **549**: 57-62
- Nakamura Y, Kaneko T, Hirose M, Miyajima N, Tabata S** (1998) CyanoBase, a www database containing the complete nucleotide sequence of the genome of Synechocystis sp. strain PCC6803. *Nucleic Acids Res* **26**: 63-67
- Nakamura Y, Kaneko T, Miyajima N, Tabata S** (1999) Extension of CyanoBase. CyanoMutants: repository of mutant information on Synechocystis sp. strain PCC6803. *Nucleic Acids Res* **27**: 66-68
- Nakamura Y, Kaneko T, Tabata S** (2000) CyanoBase, the genome database for Synechocystis sp. strain PCC6803: status for the year 2000. *Nucleic Acids Res* **28**: 72

- Nakao M, Okamoto S, Kohara M, Fujishiro T, Fujisawa T, Sato S, Tabata S, Kaneko T, Nakamura Y** (2010) CyanoBase: the cyanobacteria genome database update 2010. *Nucleic Acids Res* **38**: D379-381
- Nakashima H, Nishikawa K** (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J Mol Biol* **238**: 54-61
- Neuwald AF, Liu JS, Lawrence CE** (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci* **4**: 1618-1632
- Ng SK, Zhang Z, Tan SH, Lin K** (2003) InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res* **31**: 251-254
- Paithoonrangsarid K, Shoumskaya MA, Kanesaki Y, Satoh S, Tabata S, Los DA, Zinchenko VV, Hayashi H, Tanticharoen M, Suzuki I, Murata N** (2004) Five histidine kinases perceive osmotic stress and regulate distinct sets of genes in *Synechocystis*. *J Biol Chem* **279**: 53078-53086
- Palomeque T, Antonio Carrillo J, Munoz-Lopez M, Lorite P** (2006) Detection of a mariner-like element and a miniature inverted-repeat transposable element (MITE) associated with the heterochromatin from ants of the genus *Messor* and their possible involvement for satellite DNA evolution. *Gene* **371**: 194-205
- Park J, Lappe M, Teichmann SA** (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J Mol Biol* **307**: 929-938
- Patrick Argos MGR, Ulrich M. Grau , Herbert Zuber , Gerhard Frank , Jon Duri Tratschin** (1979) Thermal stability and protein structure. *Biochemistry* **18**: 5698-5703
- Pradeep NV, Anupama VKG, Lakshmi P** (2012) In silico characterization of Industrial important cellulases using computational tools. *Advances in life science and technology* **4**: 8-14
- Rice P, Longden I, Bleasby A** (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276-277
- Rosmarie Rippka JD, John B. Waterbury*, Michael Herdman† and Roger Y. Stanier** (1978) Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria. *Journal of General Microbiology* **111**: 1-61
- Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J** (2000) Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci U S A* **97**: 6652-6657
- Sali A, Blundell TL** (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**: 779-815
- Sasaki NV, Sato N** (2010) CyanoClust: comparative genome resources of cyanobacteria and plastids. Database (Oxford) **2010**: bap025
- Sazuka T, Ohara O** (1997) Towards a proteome project of cyanobacterium *Synechocystis* sp. strain PCC6803: linking 130 protein spots with their respective genes. *Electrophoresis* **18**: 1252-1258
- Sazuka T, Yamaguchi M, Ohara O** (1999) Cyano2Dbase updated: linkage of 234 protein spots to corresponding genes through N-terminal microsequencing. *Electrophoresis* **20**: 2160-2171
- Schwede T, Kopp J, Guex N, Peitsch MC** (2003) SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* **31**: 3381-3385
- Seino Y, Takahashi T, Hihara Y** (2009) The response regulator RpaB binds to the upstream element of photosystem I genes to work for positive regulation under low-light conditions in *Synechocystis* sp. Strain PCC 6803. *J Bacteriol* **191**: 1581-1586
- Shoumskaya MA, Paithoonrangsarid K, Kanesaki Y, Los DA, Zinchenko VV, Tanticharoen M, Suzuki I, Murata N** (2005) Identical Hik-Rre systems are involved in perception and transduction of salt signals and hyperosmotic signals but regulate the expression of individual genes to different extents in *synechocystis*. *J Biol Chem* **280**: 21531-21538

- Sinha RP, Singh SP, Hader DP** (2007) Database on mycosporines and mycosporine-like amino acids (MAAs) in fungi, cyanobacteria, macroalgae, phytoplankton and animals. *J Photochem Photobiol B* **89**: 29-35
- Sitbon E, Pietrokovski S** (2007) Occurrence of protein structure elements in conserved sequence regions. *BMC Struct Biol* **7**: 3
- Smith AA, Plazas MC** (2011) In silico characterization and homology modeling of cyanobacterial phosphoenolpyruvate carboxylase enzymes with computational tools and bioinformatics servers *American Journal of Biochemistry and molecular biology* **1**: 319-336
- Stahl PL, Lundeberg J** (2012) Toward the single-hour high-quality genome. *Annu Rev Biochem* **81**: 359-378
- Steglich C, Futschik ME, Lindell D, Voss B, Chisholm SW, Hess WR** (2008) The challenge of regulation in a minimal photoautotroph: non-coding RNAs in *Prochlorococcus*. *PLoS Genet* **4**: e1000173
- Su Z, Olman V, Xu Y** (2007) Computational prediction of Pho regulons in cyanobacteria. *BMC Genomics* **8**: 156
- Szilagyi A, Kovacs KL, Rakhely G, Zavodszky P** (2002) Homology modeling reveals the structural background of the striking difference in thermal stability between two related [NiFe]hydrogenases. *J Mol Model* **8**: 58-64
- Thompson W CS, McCue L A, Lawrence C E** (2007) Using the Gibbs Motif Sampler for phylogenetic footprinting. *Methods Mol Biol*: 395-403
- Vijayan V, Jain IH, O'Shea EK** (2011) A high resolution map of a cyanobacterial transcriptome. *Genome Biol* **12**: R47
- von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P** (2007) STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* **35**: D358-362
- Voss B, Georg J, Schon V, Ude S, Hess WR** (2009) Biocomputational prediction of non-coding RNAs in model cyanobacteria. *BMC Genomics* **10**: 123
- Wang T, Shen G, Balasubramanian R, McIntosh L, Bryant DA, Golbeck JH** (2004) The *sufR* gene (*sll0088* in *Synechocystis* sp. strain PCC 6803) functions as a repressor of the *sufBCDS* operon in iron-sulfur cluster biogenesis in cyanobacteria. *J Bacteriol* **186**: 956-967
- Wanner B** (1996) Phosphorus assimilation and control of the phosphate regulon. *Escherichia coli and Salmonella*: 1357-1381
- Wels M, Francke C, Kerkhoven R, Kleerebezem M, Siezen RJ** (2006) Predicting cis-acting elements of *Lactobacillus plantarum* by comparative genomics with different taxonomic subgroups. *Nucleic Acids Res* **34**: 1947-1958
- Whitton BA** (2012) Ecology of Cyanobacteria II. Accessed November **10**
- Wilkinson DL, Harrison RG** (1991) Predicting the solubility of recombinant proteins in *Escherichia coli*. *Biotechnology (N Y)* **9**: 443-448
- Wishart DS, Arndt D, Berjanskii M, Guo AC, Shi Y, Shrivastava S, Zhou J, Zhou Y, Lin G** (2008) PPT-DB: the protein property prediction and testing database. *Nucleic Acids Res* **36**: D222-229
- Wu J, Zhao F, Wang S, Deng G, Wang J, Bai J, Lu J, Qu J, Bao Q** (2007) cTFbase: a database for comparative genomics of transcription factors in cyanobacteria. *BMC Genomics* **8**: 104
- Zuber U, Schumann W** (1994) CIRCE, a novel heat shock element involved in regulation of heat shock operon *dnaK* of *Bacillus subtilis*. *J Bacteriol* **176**: 1359-1363

