2011 International Conference on Bioscience, Biochemistry and Bioinformatics
IPCBEE vol.5 (2011) © (2011) IACSIT Press, Singapore

Sequence Motif Analysis of Phycobiliproteins in Cyanobacterial Genome

Kuo-Yuan Hwa*
Institute of Organic and Polymeric Materials
Department of Molecular Science and Engineering
Centre for Biomedical Industries
National Taipei University of Technology
Taipei, Taiwan, ROC
Email: kyhwa@ntut.edu.tw

Govinda Lenka
Graduate Institute of Biotechnology
National Taipei University of Technology
Taipei, Taiwan, ROC

Boopathi Subramani
Institute of Organic and Polymeric Materials
National Taipei University of Technology
Taipei, Taiwan, ROC

Shu-Mei Chang
Institute of Organic and Polymeric Materials
Department of Molecular Science and Engineering
National Taipei University of Technology
Taipei, Taiwan, ROC

***Abstract*—Phycobiliproteins are a group of pigmented proteins commonly present in cyanobacteria and red algae which has potential application in different fields. Based on the color and absorbance they are grouped in to three important types. The sequence motif was analyzed for all the important Phycobiliproteins by using its protein sequence obtained from the non-redundant database. 39 sequence motifs were predicted by MEME, among them 11 were found significant with less e-value. MAST analysis showed the motifs showed high similarity with Synechococcus species and Thermococcus species in the Cyanobase.**

Keywords-Phycobiliproteins; Cyanobacteria; MEME; MAST; motif

I. INTRODUCTION

Phycobiliproteins are a group of proteins referred as light-harvesting macromolecules that function as components of the photosynthetic apparatus in cyanobacteria and several groups of eukaryotic algae [1-5]. Three major types of phycobiliproteins are phycoerythrins, phycocyanins and allophycocyanins which are present in red algae and cyanobacteria. The absorption maxima are 540–570nm for Phycoerythrin, 610–620nm for Phycocyanin, and 650–655 nm for Allophycocyanin. These proteins spans over a 200-nm portion of the visible spectrum (470–670 nm) which enable the organism to extend the range of their spectral absorption to collect light which are not absorbed completely by chlorophyll *a* and through fluorescence energy transfer, it conveys the energy to chlorophyll at the photosynthetic reaction center. The phycobiliproteins are water soluble, very stable at physiological pHs, and highly fluorescent proteins [6,7] and constitutes about 60% of the soluble protein content [8]. These are oligomeric proteins, built up from chromophore-having polypeptides belonging to two families (α and β) which may be originated from a common ancestor [4]. The phycobiliproteins are composed of a number of subunits, each having a protein backbone to which linear tetrapyrrole chromophores are covalently bound. The phycobiliproteins involved in an extremely efficient energy transfer chain in the photosynthetic reaction. These proteins transfer the excitation energy with less radiation processes to the reaction centers in the photosynthetic membranes [6].

Cyanobacteria, which produce phycobiliproteins, are recognized as an important and widespread component of marine picophytoplankton that contributes significantly to total carbon biomass and primary productivity of the oceans. The phycobiliproteins obtained from cyanobacteria have gained commercial importance, as they got several applications. The main applications of these molecules are as natural dyes, colorants in food and it is most widely as fluorescent labels for cells and macromolecules in highly sensitive fluorescence techniques [9], but various studies showed on their health-promoting properties and broad range of pharmaceutical applications [10].

In this study, the available protein sequences for phycobiliproteins from the non-redundant protein database were used to find sequence motif patterns among the cyanobacterial species. Basically, motif is a region or portion of a protein sequence that has a specific structure and is functionally significant. Protein families are often characterized by one or more such motifs. For motif prediction, MEME tool was used to analyze the sequences and produce a description (motif) for each pattern it discovers. This study was carried out to find key functional motifs among the different cyanobacterial species (Fig. 1). Detection of sequence motifs in protein is an important approach since motifs carry out and regulate various functions, and the presence of specific motifs may help to classify proteins. The motifs which are involved in regulating the synthesis of phycobiliproteins can be helpful in overproduction of the protein in commercial scale. After

analyzing the significant motifs in the cyanobacterial phycobiliproteins, it can be used to improve the properties of phycobiliproteins and its production by using various genetic engineering approaches. This study can also pave the way to produce phycobiliproteins with different absorption properties and coloring properties by altering the key residues in the significant motifs by site directed mutagenesis.

II. Materials And Methods

A. Training Dataset

The phycobiliproteins sequences retrieved from the non-redundant database, NCBI RefSeq for various phycobiliproteins. Then, these sequences were grouped into α and β subunits (Table I).

B. MEME analysis

The retrieved Sequences for the α and β subunits of the phycobiliproteins were submitted to the Multiple Em for Motif Elicitation (MEME) available at the San Diego Supercomputer Center (SDSC), (http://meme.sdsc.edu/meme4_4_0/cgibin/meme.cgi) [11] to predict the sequence motifs. In the sequence motif analysis, we have set the maximum width as 50 for each motif and Maximum number of motifs is 3. The number of occurrences of each motif was automatically chosen by MEME in order to minimize the 'E-value' of the motif-the probability of finding an equally well-conserved pattern in random sequences. The *E*-value of a motif is based on its log likelihood ratio, width, sites, the background letter frequencies, and the size of the training set.

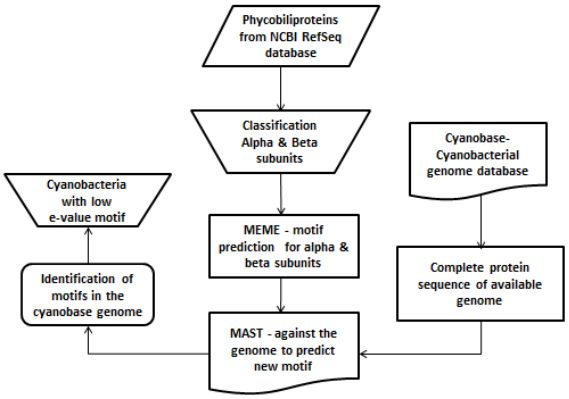


Figure 1. Phycobiliproteins sequence motif prediction work flow.

TABLE I. α And β Subunits Of Phycobiliproteins Sequences.

Pigment	α Subunit	β Subunit	Unclassified
R-Phycoerythrin	11	9	-
B-Phycoerythrin	-	-	1
C-Phycoerythrin	9	9	-
C-Phycocyanin	25	30	-
R-Phycocyanin	14	16	-
Allophycocyanin	102	148	-
Phycoerythrocyanin	9	8	-

The output of the MEME showed different motifs with the e-value less than 10. However, we have sorted the MEME output to find the common motifs with e-value below 2. MEME motifs are represented by position-specific probability matrices that specify the probability of each possible letter appearing at each possible position in an occurrence of the motif. These are displayed as 'Sequence Logos', containing stacks of letters at each position in the motif. The sequence logo of the predicted motifs was created by using Weblogo v 2.8.2 available online at Department of Plant and Microbial Biology, University of California, Berkeley, USA (http://weblogo.berkeley.edu/) [12].

C. MAST analysis

Motif Alignment and Search Tool was used to compare the MEME output against the cyanobacterial genome available in the Cyanobase. This was performed to look for other occurrences of the predicted sequence motifs in the cyanobacterial genome. In MAST analysis, non-redundant and special database were selected as the support database. MAST determines the best match in the all the sequence to each motif. The scores of the best matched sequence motifs are combined into a score for the overall match among the complete motif set and the sequence, resulting in an E-value for each sequence. The output from MAST is a list of the sequences for which the E-value less than 1.5 was specified as the threshold value. The MAST program is part of the MEME suite available at SDSC, (http://meme.sdsc.edu/meme4_4_0/cgi-bin/mast.cgi).

III. Results And Discussion

A. Prediction of motifs in phycobiliproteins

In order to find the common sequence motifs for each α and β subunits of the phycobiliproteins retrieved from the Non-redundant protein database, MEME analysis was performed. The e-values of the most common predicted motifs were listed in the table II. The MEME output produced 39 motifs with threshold e-values with less than 10, but motifs with less than 2 e-values were considered as the most significant motifs in this study. The sequence logos for the 11 motifs predicted by MEME analysis are listed in the fig. 2.

B. MAST analysis

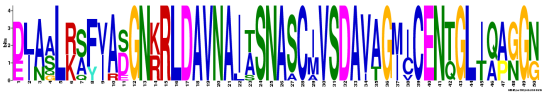
To search sequences in cyanobase for matches to the 11 motifs found by MEME, MAST program was used. The dropdown option for cyanobacterial database is currently not available in MAST. So, the motifs were searched against individual genome. The MAST output showed (Table III A and III B) similarity between the predicted sequence motifs and the cyanobacterial genome with an e-value of 10. Table III A and III B summarized the top scoring sequences with a sequence E-value better than the threshold. In this analysis, the best possible matches for the 11 motifs were found in 25

cyanobacterial genome. The sequences with e-values less than 1.5 were found to be having high similarity with the motifs. The motifs for the different phycobiliproteins pigments showed high similarity with the target genome sequences especially with the strains of *Synechococcus* species and *Thermococcus* species.

TABLE II. PHYCOBILIPROTEINS MOTIFS PREDICTED BY MEME. THE E-VALUES WITH LESS THAN 2 ARE MARKED IN BLUE.

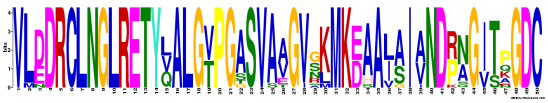
Pigment	Subunit classification	No. of motifs	e-value
C-Phycoerythrin	Alpha	1	7.5e-335
		2	8.9e-269
		3	3.6e-170
	Beta	1	2.7e-250
		2	**1.5e-193**
		3	2.6e-112
C-Phycocyanin	Alpha	1	**1.3e-506**
		2	6.3e-704
		3	5.5e-325
	Beta	1	3.8e-910
		2	**1.1e-966**
		3	3.6e-822
Phycoerythrocyanin	Alpha	1	2.5e-388
		2	3.9e-295
		3	2.4e-293
	Beta	1	**1.8e-269**
		2	2.9e-214
		3	**1.5e-178**
R-Phycoerythrin	Alpha	1	**1.9e-327**
		2	3.1e-259
		3	6.5e-108
	Beta	1	9.8e-248
		2	**1.3e-155**
		3	2.8e-084
R-Phycocyanin	Alpha	1	5.6e-399
		2	3.5e-288
		3	**1.9e-023**
	Beta	1	**1.2e-499**
		2	2.5e-108
		3	5.0e-094
Allophycocyanin	Alpha	1	4.8e-415
		2	**2.0e-333**
		3	9.8e-316
	Beta	1	4.4e-450
		2	4.9e-348
		3	6.3e-331
B-Phycoerythrin		1	3.1e-084
		2	3.5e-029
		3	**1.3e-023**

C Phycoerythrin Beta subunit e-value: 1.5e-193



R Phycocyanin Alpha subunit e-value: 1.9e-023

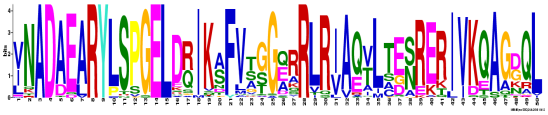
R Phycocyanin Beta subunit e-value: 1.2e-499



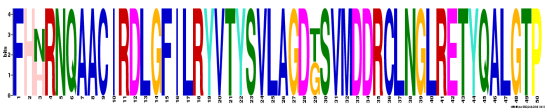
C Phycocyanin Alpha subunit e-value: 1.3e-506

C Phycocyanin Beta subunit e-value: 1.3e-506

Allophycocyanin Alpha subunit e-value: 2.0e-3334



Phycoerythrocyanin Beta subunit 1 e-value: 1.8e-269



Phycoerythrocyanin Beta subunit 2 e-value: 1.8e-269

B Phycoerythrin e-value: 1.3e-023

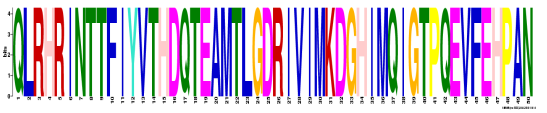


Figure 2. The sequence logos of the 11 motifs predicted by MEME. The total height of the stack is the "information content" of that position in the motif in bits. The height of a letter indicates its relative frequency at the given position (x-axis) in the motif.

TABLE III. A. MAST ANALYSIS FOR BEST POSSIBLE MATCHED SEQUENCES E-VALUE LESS THAN 10 FOR EACH CYANOBACTERIAL GENOME. THE E-VALUES BELOW THE THRESHOLD (1.5) WERE MARKED IN RED.

Cyanobacterial genome	R-PHYCOERYTHRIN		C-PHYCO ERYTHRIN	B PHYCOER YTHRIN	R-PHYCOCYANIN		C-PHYCOCYANIN	
	α Subunit	β Subunit	β Subunit	MOTIF 3 e-value	α Subunit	β Subunit	α Subunit	β Subunit
	MOTIF 1 e-value	MOTIF 2 e-value	MOTIF 2 e-value		MOTIF 3 e-value	MOTIF 1 e-value	MOTIF 1 e-value	MOTIF 2 e-value
Chlorobium tepidum TLS	1.4e-05	4.9e-05	4.7e-05	1.0e-07	1.3e-05	4.2e-06	1.2e-06	-
Prochlorococcus marinus MED4	1.0e-06	1.2e-05	1.3e-05	1.8e-11	1.7e-06	1.1e-05	2.0e-07	-
Prochlorococcus marinus MT9313	1.4e-06	1.0e-05	1.1e-05	1.0e-22	2.7e-05	2.0e-07	1.3e-05	-
Prochlorococcus marinus SS120	1.7e-05	1.7e-05	2.2e-05	1.2e-06	1.1e-06	1.4e-06	1.5e-42	-
Prochlorococcus marinus str. AS9601	1.7e-05	1.3e-05	1.2e-05	1.4e-13	1.8e-06	1.1e-05	2.1e-06	-
Prochlorococcus marinus str. MIT 9211	1.1e-08	1.3e-50	1.0e-06	1.3e-13	1.6e-16	1.0e-05	1.4e-05	-
Prochlorococcus marinus str. MIT 9215	1.3e-05	8.4e-05	1.5e-06	1.3e-13	1.3e-05	2.3e-06	1.0e-05	-
Prochlorococcus marinus str. MIT 9301	1.3e-06	9.1e-07	1.2e-05	1.5e-09	1.8e-05	1.6e-05	2.1e-06	-
Prochlorococcus marinus str. MIT 9303	1.4e-06	1.0e-05	1.2e-33	1.3e-12	2.8e-05	2.1e-06	3.4e-42	-
Prochlorococcus marinus str. MIT 9312	8.3e-06	1.2e-06	1.2e-05	1.4e-14	1.7e-05	1.8e-05	2.4e-05	-
Prochlorococcus marinus str. MIT 9515	1.1e-06	2.0e-06	1.5e-05	1.1e-11	1.2e-05	1.4e-06	1.2e-05	-
Prochlorococcus marinus str. NATL1A	1.2e-05	1.3e-05	1.4e-08	2.0e-05	1.3e-16	1.2e-05	1.1e-40	-
Prochlorococcus marinus str. NATL2A	4.0e-06	1.2e-05	2.5e-34	1.5e-05	6.10e-31	1.0e-19	1.1e-40	-
Synechococcus sp. CC9605	1.0e-16	1.1e-06	1.2e-06	1.0e-05	1.4e-57	1.1e-05	1.0e-53	-
Synechococcus sp. CC9902	1.4e-11	1.3e-06	1.0e-52	1.3e-07	1.3e-0.5	1.1e-05	1.0e-53	-
Synechococcus sp. JA-2-3Ba(2-13)	1.0e-14	1.1e-05	1.1e-09	1.3e-07	1.2e-05	1.1e-05	1.1e-05	-
Synechococcus sp. JA-3-3Ab	1.0e-14	1.2e-27	1.4e-24	1.1e-06	1.3e-35	1.4e-10	1.1e-06	-
Synechococcus sp. PCC 7002	1.0e-06	1.6e-07	1.2e-17	1.1e-06	1.5e-05	1.0e-11	1.7e-05	-
Synechococcus WH8102	1.3e-49	1.0e-06	1.0e-05	1.1e-14	2.3e-39	1.2e-05	1.0e-53	-
Synechococcus CC9311	1.4e-11	1.1e-05	1.5e-05	1.3e-07	1.4e-06	1.5e-05	1.0e-53	-
Synechococcus elongatus PCC 6301	1.4e-18	1.3e-05	1.0e-05	1.0e-06	2.5e-39	2.3e-07	1.5e-24	-
Synechococcus elongatus PCC 7942	1.2e-37	1.4e-06	1.8e-05	1.0e-06	2.5e-39	2.3e-07	1.2e-06	-
Synechococcus sp. RCC307	1.2e-07	1.0e-05	1.0e-04	1.3e-07	1.2e-05	1.7e-05	1.0e-53	-
Synechococcus sp. WH 7803	1.1e-05	1.5e-05	1.2e-05	1.1e-08	1.3e-05	1.1e-05	1.2e-05	-
Thermococcus sp.	1.1e-05	1.3e-07	1.2e-06	1.1e-08	1.1e-06	1.0e-07	1.2e-05	-

TABLE III. B. MAST ANALYSIS FOR BEST POSSIBLE MATCHED SEQUENCES WITH E-VALUE LESS THAN 10 FOR EACH CYANOBACTERIAL GENOME. THE E-VALUES BELOW THE THRESHOLD (1.5) WERE MARKED IN RED.

Cyanobacterial genome	ALLOPHYCOCYANIN	PHYCOERYTHROCYANIN	
	α Subunit	β Subunit	
	MOTIF 2 e-value	MOTIF 1 e-value	MOTIF 2 e-value
Chlorobium tepidum TLS	1.0e-05	3.3e-05	-
Prochlorococcus marinus MED4	1.9e-07	1.0e-14	2.1e-06
Prochlorococcus marinus MT9313	1.2e-07	2.1e-05	2.2e-05
Prochlorococcus marinus SS120	4.0e-05	2.3e-05	1.1e-05
Prochlorococcus marinus str. AS9601	1.4e-06	7.3e-05	1.5e-05
Prochlorococcus marinus str. MIT 9211	1.8e-05	1.4e-10	1.5e-05
Prochlorococcus marinus str. MIT 9215	1.2e-05	4.4e-05	1.1e-06
Prochlorococcus marinus str. MIT 9301	1.2e-06	2.4e-05	1.5e-05
Prochlorococcus marinus str. MIT 9303	1.2e-06	1.9e-05	1.0e-05
Prochlorococcus marinus str. MIT 9312	1.2e-06	1.6e-05	1.1e-05
Prochlorococcus marinus str. MIT 9515	1.0e-06	1.0e-05	1.5e-05
Prochlorococcus marinus str. NATL1A	2.6e-06	1.1e-10	1.4e-06
Prochlorococcus marinus str. NATL2A	5.1e-06	1.1e-05	2.6e-09
Synechococcus sp. CC9605	1.0e-11	1.0e-05	1.3e-12
Synechococcus sp. CC9902	1.0e-05	1.1e-05	1.1e-06
Synechococcus sp. JA-2-3Ba(2-13)	1.2e-10	1.0e-07	1.1e-05
Synechococcus sp. JA-3-3Ab	1.1e-08	1.2e-13	1.1e-05
Synechococcus sp. PCC 7002	1.4e-05	1.0e-21	1.2e-05
Synechococcus WH8102	1.6e-05	1.1e-17	1.1e-13
Synechococcus CC9311	1.1e-06	1.2e-06	1.1e-21
Synechococcus elongatus PCC 6301	1.0e-15	1.5e-17	1.1e-06
Synechococcus elongatus PCC 7942	1.0e-15	1.5e-17	1.1e-06
Synechococcus sp. RCC307	1.0e-33	1.1e-34	1.2e-13
Synechococcus sp. WH 7803	1.1e-10	1.2e-05	1.5e-09
Thermococcus sp.	1.1e-05	1.0e-06	1.1e-06

IV. CONCLUSION

In further continuation of this work, we have planned to investigate the functional activities of each of the key motifs found in our study. At present, the online tool available only for few well characterized genomes for annotating the predicted sequence motifs, and the tool for annotating the motifs in cyanobacterial genome is yet to be developed which also depends on the characterization of the cyanobacterial genome. In addition, we would like to find the statistically significant associations between significant motifs found in our study and functional gene sets, in order to understand the biological roles of transcription factors. Performing the recombinant DNA technology methods including cloning and site directed mutagenesis will give promising information on the properties and functional activities of phycobiliproteins which further can help the researcher to enhance the quality and quantity of these proteins in cyanobacterial species. Transfection and expression of recombinant DNA of these motifs can be used to study the effects and regulation of genes and their encoded proteins.

ACKNOWLEDGEMENT

The research is supported by the joint research fund of TMU-NTUT and research grant NTUT 99-140-04 to KuoYuan Hwa. The support for Boopathi Subramani is from the Energy and Optoelectronics Materials Program at the National Taipei University of Technology and The Centre for Biomedical Industries.

REFERENCES

[1] R. MacColl, "Cyanobacterial Phycobilisomes," Journal of Structural Biology 124, 1998, pp.311-334

[2] H. Zuber, "The structure of light-harvesting pigment-protein complexes," In The Light Reactions (Barber, J., ed.), 1987, pp. 197-259. Elsevier Biomedical, Amsterdam.

[3] A. N. Glazer, "Phycobilisomes: assembly and attachment," In The Cyanobacteria (Fay, P. & Van Baalen, C., eds.), 1987, pp.69-94. Elsevier Biomedical, Amsterdam.

[4] A. N. Glazer, "Light guides. Directional energy transfer in a photosynthetic antenna," J. Biol. Chem. vol. 264, 1989, pp.1-4.

[5] N. Tandeau de Marsac, D. Mazel, V. Capuano, T. Damerval, and J. Houmard, "Genetic analysis of the cyanobacterial light-harvesting antenna complex," In Molecular Biology of Membrane-bound Complexes in Phototrophic Bacteria (Drews, G. & Daves, E. A., eds), 1990, pp. 143-154. Plenum Press, New York.

[6] A.N. Glazer, "Chemicals from Microalgae," in: Z. Cohen (Ed.), Taylor & Francis, Philadelphia, 1999, pp. 263.

[7] A.N. Glazer, "Phycobiliproteins-a family of valuable, widely used fluorophores, J. Appl. Phycol. vol. 6, 1994, pp. 105-112.

[8] L. Bogorad, "Phycobiliproteins and complimentary chromatic adaptation," Annu. Rev. Plant Physiol. vol. 26, 1975, pp. 369-401.

[9] A.N. Glazer and L. Stryer, Phycofluor probes, Trends Biochem. Sci. vol. 9, 1984, pp. 423–427

[10] S. Sekar and M. Chandramohan, "Phycobiliproteins as a commodity: trends in applied research, patents and commercialization," J. Appl. Phycol. vol. 20, 2008, pp.113–136.

[11] T.L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, California, 1994, pp. 28-36.

[12] G.E. Crooks, G. Hon, J.M. Chandonia and S.E. Brenner, "WebLogo: A sequence logo generator," Genome Research, vol. 14, 2004, pp. 1188-1190.