CHAPTER 3 MATERIALS AND METHODS (I)

3.1 Meterials

3.1.1 Cyanobacterial genomic data

The 39 completely sequenced and 13 draft sequenced cyanobacterial genomic data were obtained from the NCBI database (http://www.ncbi.nlm.nih.gov/). The characteristics of 49 cyanobacterial genomes used in this study and their general genomic properties are shown in the table 3.1.

3.1.2 Computer resources

1) CPU: Intel(R) Core(TM) 2 Duo CPU @ 2.20GHz
2) Hard disk: 250 GB
3) Memory: 2 GB

3.1.3 Programming languages

The algorithms used for reconstructing the parsimonious evolutionary scenarios were coded in form of Python Programming language. In addition, the source codes were interpreted underneath the computer environment of Python 2.6 incorporating with biopython package.

3.1.4 Bioinformatics software and tools

For assigning the homology of protein sequences, the NCBI standalone BLAST was used. The OrthoMCL tool was applied to construct the cluster of cyanobacterial protein genes. The evolutionary tree reconstruction was used the MEGA software version 4 (Tamura, et al., 2007). Moreover, the SplitTree 4.8 program (Huson and Bryant, 2006) was used for constructing phylogenetic networks of the phyletic pattern of cyanobacterial genes.

Table 3.1 The cyanobacterial genomes have been included in this study and its general features.

Genome	Abbreviation	Size (Mb)*	Protein coding gene	RNA coding Gene	%(G+C)	project ID
Acaryochloris marina MBIC11017 MBIC 11017	AMA1017	8.36	8383	79	47.3	12997
Arthrospira maxima CS-328[§]	AMAS328	6.00	5690	38	44.8	29085
Anabaena variabilis ATCC 29413	AVA9413	7.07	5661	63	41.4	10642
Cyanothece sp. ATCC 51142	CAT1142	5.43	5304	55	37.0	20319
Cyanothece sp. CCY0110[§]	CCC1110	5.90	6475	45	36.7	18951
Cyanothece sp. PCC 7424	CPC7424	6.52	5710	57	38.6	20479
Cyanothece sp. PCC 7425	CPC7425	5.82	5327	57	50.8	28337
Cyanothece sp. PCC 7822[§]	CPC7822	5.70	5227	39	40.0	28535
Cyanothece sp. PCC 8801	CPC8801	4.81	4367	53	39.8	20503
Cyanothece sp. PCC 8802[§]	CPC8802	4.70	4642	39	39.8	28339
Crocosphaera watsonii WH 8501[§]	CWA8501	6.24	5958	38	37.1	10651
Gloeobacter violaceus PCC 7421	GVI7421	4.66	4430	52	62.0	9606
Lyngbya sp. PCC 8106 PCC8106[§]	LPC8106	7.00	6142	43	41.1	13409
Microcystis aeruginosa NIES-843	MEAN843	5.80	6312	52	42.3	27835
Nostoc sp. PCC 7120	NPC7102	7.71	6130	83	41.3	244
Nostoc punctiforme PCC 73102	NPU3102	9.01	6690	104	41.4	216
Nodularia spumigena CCY9414[§]	NSP9441	5.30	4860	44	41.3	13447
Prochlorococcus marinus subsp. marinus str. CCMP1375	PMA1375	1.75	1883	46	36.4	419
Prochlorococcus marinus subsp. pastoris str. CCMP1986	PMA1986	1.70	1717	44	30.8	213
Prochlorococcus marinus str. MIT 9211	PMA9211	1.70	1855	45	38.0	13551
Prochlorococcus marinus str. MIT 9215	PMA9215	1.70	1983	42	31.1	18633
Prochlorococcus marinus str. MIT 9301	PMA9301	1.60	1907	42	31.3	15746
Prochlorococcus marinus str. MIT 9303	PMA9303	2.70	2997	52	50.0	13496
Prochlorococcus marinus str. MIT 9312	PMA9312	1.70	1810	45	31.2	13910
Prochlorococcus marinus str. MIT 9313	PMA9313	2.41	2269	55	50.7	220

Table 3.2 The cyanobacterial genomes have been included in this study and its general features (continue).

Prochlorococcus marinus str. MIT 9515	PMA9315	1.70	1906	42	30.8	13617
Prochlorococcus marinus str. AS9601	PMA9601	1.70	1921	43	31.3	13548
Prochlorococcus marinus str. NATL1A	PMANA1A	1.90	2193	43	35.0	15660
Prochlorococcus marinus str. NATL2A	PMANA2A	1.80	2163	44	35.1	13911
Synechococcus sp. CC9311	SCC9311	2.61	2892	52	52.4	12530
Synechococcus sp. CC9902	SCC9502	2.20	2307	51	54.2	13655
Synechococcus sp. CC9605	SCC9605	2.51	2645	54	59.2	13643
Synechococcus elongatus PCC 6301	SEL6301	2.70	2527	55	55.5	13282
*Synechococcus elong*atus PCC 7942	SEL7942	2.75	2662	53	55.5	10645
Synechococcus sp. JA-3-3Ab A-Prime	SJA23BA	2.90	2760	55	60.2	16251
Synechococcus sp. JA-2-3B'a(2-13) B-Prime	SJA33AB	3.00	2862	52	58.5	16252
Synechocystis sp. PCC 6803	SPC6803	4.00	3569	50	47.7	60
Synechococcus sp. PCC 7002	SPC7002	3.40	3186	49	49.6	28247
Spirulina platensis C1[§]	SPLATC1	5.80	6360	40	44.6	-
Synechococcus sp. RCC307	SRCC307	2.20	2535	48	60.8	13654
Synechococcus sp. RS9916[§]	SRS9916	2.66	2961	48	59.8	13557
Synechococcus sp. RS9917[§]	SRS9917	2.58	2770	50	64.5	13555
Synechococcus sp. WH 5701[§]	SWH5701	3.04	3346	55	65.4	13554
Synechococcus sp. WH 7803	SWH7803	2.40	2533	53	60.2	13642
Synechococcus sp. WH 7805[§]	SWH7805	2.62	2883	51	57.6	13553
Synechococcus sp. WH 8102	SWH8102	2.43	2519	55	59.4	230
Synechococcus sp. BL107[§]	SYBL107	2.30	2507	46	54.2	13559
Thermosynechococcus elongatus BP-1	TELOBP1	2.59	2476	49	53.9	303
Trichodesmium erythraeum IMS101	TERI101	7.80	4451	50	34.1	318

*Include plasmid if it available, [§]Draft sequenced genome

3.2 Methodologies

3.2.1 Overview

The aim of this study is to uncover the cyanobacterial evolutionary scenarios by using the available public genomic sequences. Overall methodologies of this study can be briefly described as a flowchart in a figure 3.1. From the collection of all cyanobacterial proteomes which are available in the NCBI database, the cyanoCOGs construction was performed by using the OrthoMCL methods. The resulted cyanoCOGs were used to identify the phyletic pattern, and reconstruct the cyanobacterial evolutionary tree. Then, the phyletic pattern used to determine the evolutionary scenario of each individual protein according to the reconstructed evolutionary tree. Finally, the whole gene gain and loss event in cyanobacterial lineages could be uncovered.

3.2.2 Construction of cyanobacterial clusters of orthologous groups of proteins (cyanoCOGs)

OrthoMCL procedure starts with all-against-all BLASTP comparisons of a set of protein sequences from genomes of interest as show in figure 3.2. Putative orthologous relationships are identified between pairs of genomes by reciprocal best similarity pairs. For each putative ortholog, probable "recent" paralogs are identified as sequences within the same genome that are (reciprocally) more similar to each other than either is to any sequence from another genome. Based on empirical studies, a P-value cut-off of $1 \times e^{-5}$ was chosen for putative orthologs or paralogs.

Following, putative orthologous and paralogous relationships are converted into a graph in which the nodes represent protein sequences, and the weighted edges represent their relationships. As shown in figure 3.3, weights are initially computed as the average-Log(P-value) of BLAST results for each pair of sequences. Because the high similarity of "recent" paralogs relative to orthologs can bias the clustering process, edge weights are then normalized to reflect the average weight for all ortholog pairs in these two species (or "recent" paralogs when comparing within species). Although more sophisticated weighting schemes can be envisioned, this simple method for adjusting the systematic bias between edges connecting sequences within the same genome and edges connecting sequences from different genomes seems to generate satisfactory results, judging from the comparison with INPARANOID, the EGO database, and EC annotations. The resulting graph is represented by a symmetric similarity matrix to which the MCL algorithm is applied. MCL uses flow simulation and considers all the relationships in the graph globally and simultaneously during clustering, providing a robust method for separating diverged paralogs, distant orthologs mistakenly assigned based on (weak) reciprocal best hits, and sequences with different domain structures. An important parameter in the MCL algorithm is the inflation value, regulating the cluster tightness (granularity); increasing the inflation value increases cluster tightness (see below). Clusters containing sequences from at least two species form the final output of this procedure: clustered groups of orthologs and "recent" paralogs.

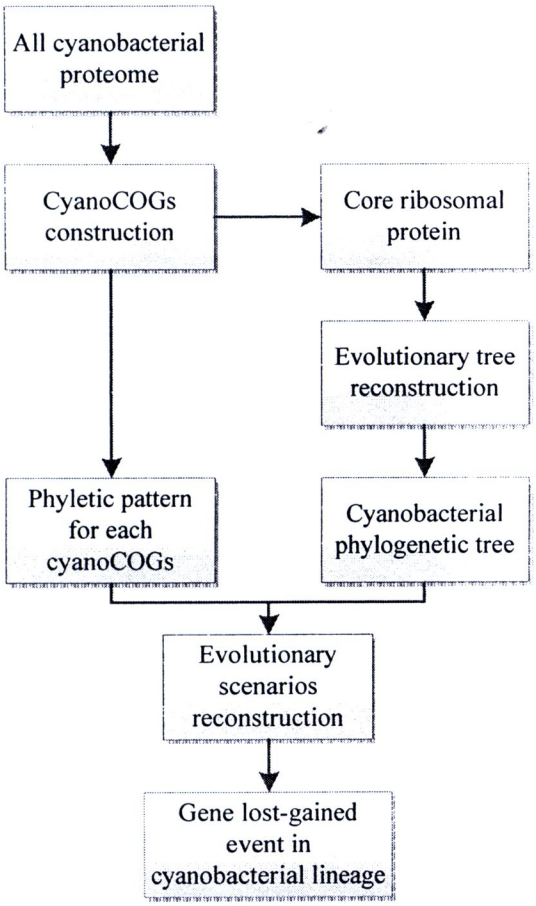


Figure 3.1 Overall methodologies for finding the evolutionary scenario of cyanobacterial genes.

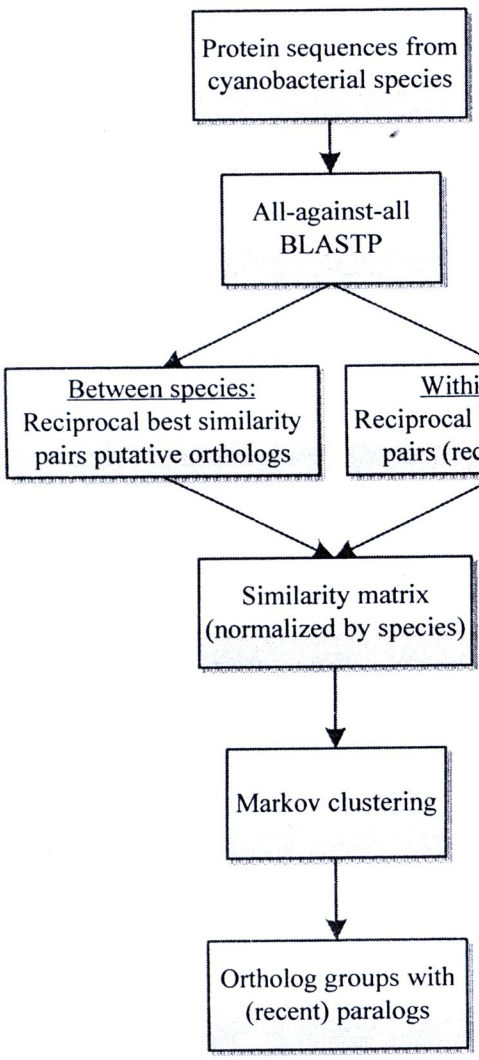
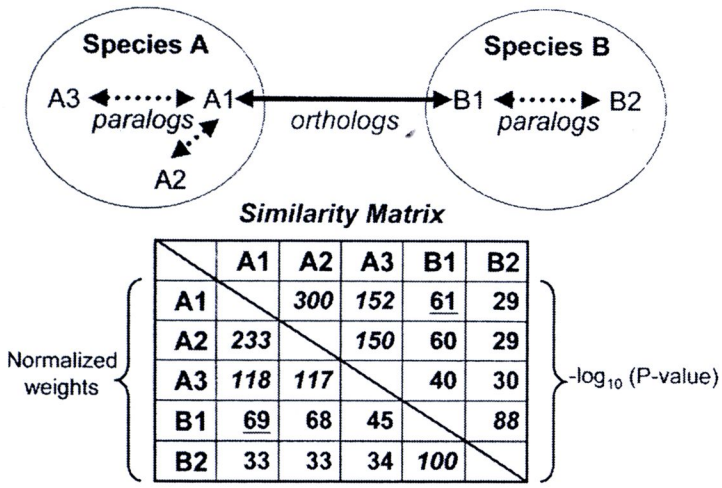


Figure 3.2 Flow of the OrthoMCL algorithm for find orthologous group of proteins (Li, *et al.*, 2003).



	A1	A2	A3	B1	B2
A1		*300*	*152*	61	29
A2	*233*		*150*	60	29
A3	*118*	*117*		40	30
B1	69	68	45		*88*
B2	33	33	34	*100*	

Figure 3.3 Illustration of sequence relationships and similarity matrix construction (Li, *et al.*, 2003).

The ortholog (originated by specification) are represented by the solid arrows, and the "recent" paralog are represented by the dotted arrows (originated by genetic duplication after specification). The upper right half of the matrix contains initial weights calculated as average $-\log_{10}(\text{p-value})$ from pairwise BLASTP similarities. The net result of the normalization by an average weight among all orthologous pairs, is to correct for systematic differences in comparisons between two species (e.g., differences attributable to nucleotide composition bias), and within a same species, to minimize the impact of "recent" paralogs (duplication within a given species) on the clustering of cross-species orthologs (Li, *et al.*, 2003).

3.2.3 Cyanobacterial lineage reconstruction

The cyanobacterial evolutionary tree was reconstructed by using the concatenated ribosomal proteins as shown in the figure 3.4. A straightforward phylogenetic analysis consists of four following steps, which every step were done by MEGA 4.1 program (Tamura, *et al.*, 2007). In addition, the phylogenetic trees of the photosynthetic apparatus genes have also been reconstructed with this method.

1) A typical alignment procedure involves the application of a program such as CLUSTALW, followed by manual alignment editing and submission to a tree building program.
2) The substitution model should be given the same emphasis as alignment and tree building. For protein sequences from closely related species BLOSUM80 substitution matrix are appropriated for determining the substitution model.
3) Neighbor Joining (NJ) was performed for building the evolutionary tree of cyanobacteria.
4) The 1,000 bootstrapping, which mean the 1,000 phylogenetic trees that reconstructed from the resampling the protein sequences, was performed in order to evaluate the robustness of the reconstructed phylogenetic tree.

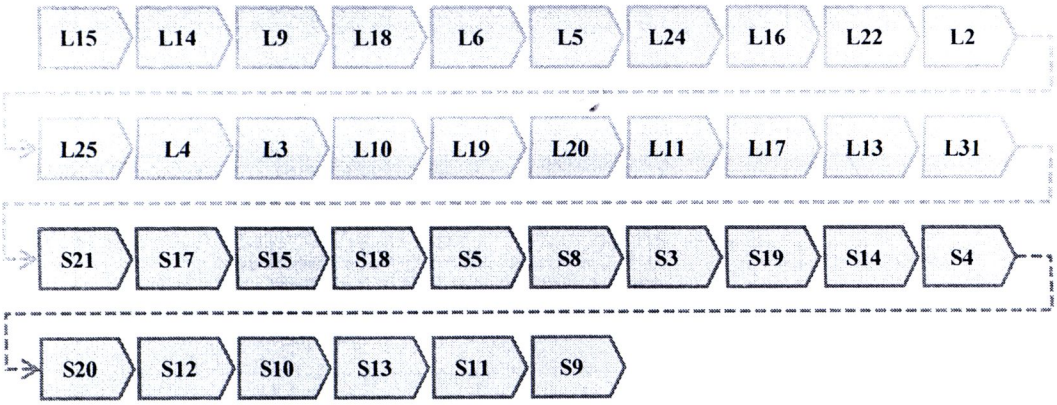


Figure 3.4 The concatenated ribosomal proteins are used for reconstructing the inferred evolution of cyanobacterial lineage. The label 'L' and 'S' represent the large and small ribosomal protein subunit, respectively.

3.2.4 Evolutionary scenarios reconstruction

The problem of building a parsimonious evolutionary scenario, given a gene's phyletic pattern and a binary evolutionary tree, can be formalized in term of bottom-up fashion. This is achieved by building a parsimonious scenario for a parent given parsimonious scenario for its children. This required maintaining, at each node of the tree, set of loss and gain events under both the assumption that the gene has been inherited at the node (A_i) and the assumption that it has not been inherited (A_n).

Let now consider the parent-children triple show in the figure 3.5 Each node in the triple is assigned with the set of loss and gain events under each of the above inheritance assumptions: $\begin{bmatrix} G_i & L_i \\ G_n & L_n \end{bmatrix}$ for the parent and similar quadruples for the children (see figure 3.5). The set G_i refers gain events in the subtree descending from the parent assumption A_i. Conversely, the set L_i contains the nodes, which have been lost under assumption A_i. The set G_n and L_n have similar meaning, but under the non-inheritance assumption A_n. Let denote the total number of events by $e_i = |G_i| + |L_i|$ under A_i, and by $e_n = |G_n| + |L_n|$ under A_n. These will be referred to the i-inconsistency and n-inconsistency of the given node, respectively. An evolution scenario, at given node, is thus defined by a pair of sets (G, L) representing the gains and losses in the sub-tree rooted at the node and use (G_i, L_i) and (G_n, L_n) to denote scenarios under assumptions A_i and A_n, respectively.

How can these sets in the parent be derived from those in the children? First, under assumption A_i, the sets G_i and L_i given all the loss and gain sets at the children will be determined. There are two alternative scenarios: (*i*) the gene has been lost in the parent, or (*ii*) the gene has not been lost in the parent. In the first case, the lost gene could not have been inherited by the children and, thus, sets L_{n1} and L_{n2} are the relevant loss events and sets G_{n1} and G_{n2} are the relevant gain events. The sets for the parent are then determined by combining the corresponding set for the children:

$$\begin{aligned} G_i &= G_{n1} \cup G_{n2} \\ L_i &= L_{n1} \cup L_{n2} \cup \{\text{parent}\} \end{aligned} \qquad (1)$$

The parent is added in the latter equation because of the assumed loss event. In the second case, the gene has been inherited and not lost; thus the loss/gain event sets will be determined by the other sets of events in the children, viz. L_{i1}, L_{i2}, G_{i1}, and G_{i2}. The sets at the parent are given by:

$$\begin{aligned} G_i &= G_{i1} \cup G_{i2} \\ L_i &= L_{i1} \cup L_{i2} \end{aligned} \qquad (2)$$

PARENT	Gains	Losses
Inherited	G_i	L_i
Not inherited	G_n	L_n

CHILD1	Gains	Losses
Inherited	G_{i1}	L_{i1}
Not inherited	G_{n1}	L_{n1}

CHILD2	Gains	Losses
Inherited	G_{i2}	L_{i2}
Not inherited	G_{n2}	L_{n2}

Figure 3.5 Patterns of events in a parent-children triple according to a parsimonious scenario (Mirkin, *et al.*, 2003).

Of the two alternatives, principle of parsimonious suggests selecting the one with the smaller number of events. Under scenario (*i*), the total number of events is $e_i = e_{n1} + e_{n2} + 1$ and, under scenario (*ii*), the total is $e_i = e_{i1} + e_{i2}$, according to (3) and (4), respectively. Parsimony suggests to selecting the minimal total score scenario:

$$e_i = \min\left(e_{n1} + e_{n2} + 1, \quad e_{i1} + e_{i2}\right) \tag{3}$$

When $e_{n1} + e_{n2} + 1 = e_{i1} + e_{i2}$ either scenario may be selected. This ambiguity is removed by using external criterion. For example, scenario (ii) may be preferred because this does not introduce additional events in the parent.

Let now determine the sets of G_n and L_n under the assumption A_n. There are again two alternative scenarios: (*i*) the gene has been gain in the parent, or (*ii*) the gene has not been gain in the parent. In the first case, the gain gene should be inherited by the childrenand, thus, to determine G_n and L_n sets L_{i1} and L_{i2} are the relevant loss events, and sets G_{i1} and G_{i2} are the relevant gain events. Then:

$$\begin{aligned} G_n &= G_{i1} \cup G_{i2} \cup \{\text{parent}\} \\ L_n &= L_{i1} \cup L_{i2} \end{aligned} \tag{4}$$

The parent is added in the former equation because of the assumed gain events.

Under scenario (*ii*), the gene has not been gained; thus, the loss and gain event sets will be determined by the other sets at the children, which yield:

$$\begin{aligned} G_n &= G_{n1} \cup G_{n2} \\ L_n &= L_{n1} \cup L_{n2} \end{aligned} \tag{5}$$

Parsimony requires that the scenario with the smaller number of events is selected. The total number of events is $e_n = e_{i1} + e_{i2} + 1$ under scenario (*i*) and $e_n = e_{n1} + e_{n2}$ under scenario (*ii*), according to (6) and (7), respectively. As discussed above, the likelihood of gains and losses may not be equal; losses are generally considered to be more likely than gain. Therefore gains may be charged with a penalty, *g*, which corresponds to the generalized parsimony approach. Taking this into account, e_i and e_n were redefined as:

$$\begin{aligned} e_i &= g \cdot |G_i| + |L_i| \\ e_n &= g \cdot |G_n| + |L_n| \end{aligned}$$

This modify the recurrence under scenario (*i*) to $e_n = e_{i1} + e_{i2} + g$. Thus, the scenario to be selected is defined by:

$$e_n = \min\left(e_{i1} + e_{i2} + g, \quad e_{n1} + e_{n2}\right) \tag{6}$$

When $e_{n1} + e_{n2} + g = e_{i1} + e_{i2}$, we may once again remove the ambiguity by selecting the scenario according to external criterion. For instance, scenario (*ii*) may be preferred as it introduces no additional gain events at the parent.