CHAPTER 2 LITERATURE REVIEWS (I)

Evolutionary study is a fundamental key step to uncover the diversities and complexities of organisms. In order to gain better understanding complexities of the evolutionary dynamics, the huge amount of the available genome sequences on the public databases is an intelligent living fossil, and the comparative genomics is powerful equipment for the evolutionary study. However, homology detection strategies are challenges aspects to identify the genes that originate from the same source across various genomes. Then, the backgrounds, methodologies, advantages, and disadvantages of each homology detection strategies are well described in this chapter. In particular, the comparative genomic studiesin the evolutionary, phylogenetic, and phylogenomic point of view would also be clarified and stated the previous knowledge and works(Eisen and Fraser, 2003).

Cyanobacteria – one of the earliest branching groups of organisms on this planet, also, the only known prokaryotes to carry out oxygenic photosynthesis (Bekker, et al., 2004) – are interesting organism for performing the comparative genomic analysis, evolutionary, and phylogenomic study. Despite their morphologic and genomic evolution, environmental niches, and biogeochemistry have been widely well-studied, the evolution of photosynthetic apparatus remain ambiguous (Bhaya, 2004). In order to reveal the cyanobacterial genomic evolution and delineate their evolutionary scenarios along the cyanobacterial lineages, the public genome sequence data of these organismal genera have provided an opportunity for a new episode in the community of cyanobacterial research. Furthermore, the knowledge of their photosynthesis machineries and processes and their evolutionary studies in this vulnerable organism are reviewed in this chapter.

Here, this chapter specifically aims to put out the comprehensive knowledge of the comparative genomic studies, including the homology detection strategies, and the phylogenetic analysis. Also, the cyanobacteria and their photosynthesis machineries, and the previous knowledge of their evolutionary are also delineated in this chapter in order to gain the better understanding of previous work and easily to understand this work.

2.1 Comparative genomic study

Comparative genomic is use for inferring the genomic functionally from one genome, which have been studied before to another one with newly sequenced. Comparative genomic was firstly introduced for studying the biology of bacteria when there were only two bacterial genomes (Fleischmann, *et al.*, 1995). The comparative genomic plays an important role on the protein functional prediction assignment, which based on the conservative protein sequences along several genomes(Huynen, *et al.*, 2004). Therefore, the comparative analyses of genome sequences are a major part of finding a functional part of the DNA sequences (Hardison, 2003). Not only a functional assignment for DNA sequences,the comparative genomics can also be used as powerful tools for revealing the diversity and tracing the evolutionary of the genome (Koonin, *et al.*, 2000). Moreover, the comparative genomics can be used for determining the minimal gene sets of the interested organisms and the gene sets of the every ancestral form (Koonin, 2003). In this genomic era with an exponential growth of genome sequences, comparative genomics have been

proofed as efficient equipment for revealing information, which have been stored inside the genomic sequences and inherited by generation to generation.

2.1.1 Homology detection strategies

Identification of homology in the genome sequences is a critically important aspect for the comparative genomics. The homology detection is also regularly used in genome annotation, gene function characterization, evolutionary genomics, and in the identification of conserved regulatory elements. In this field, the homology consists of orthology and paralogy. The orthology sequences are the homology sequences, which were separated by the speciation processes. On the other hand, the paralogy sequences are the homology sequences, which were duplicated inside the same organism. The concepts of orthology and paralogy are well-established in classical and molecular systematics, and have been extended to describe the more complicated situations, that are associated with extensive gene duplications commonly observed in eukaryotic species. In- and out-paralogs are analogous to the phylogenetic concepts in- and out-groups, denoting genes duplicated subsequent or prior to speciation, respectively. Recent duplications yield in-paralogs that may exhibit a many-to-one or many-to-many ortholog relationship with genes in the other species (termed co-orthologs) (Chen, et al., 2007).

The strategies to detect the orthologous gene are different for both conceptual and practical ways that illustrated in the figure 2.1 and summarized in table 2.1. With based on a grouping methodology, the orthologous gene detection strategies can be separated as following:

1) Tree based methods (based on the gene phylogenetic tree)
2) Network or graph based methods (based on graph)
3) Hybrid methods (combine both of the tree and graph based methods)

The tree based methods start with the collection of homologous sequences. A multiple sequence alignments, and reconstruction phylogenetic tree are performed. Then, the relationships can be analyzed either in the present or absent of 'known' phylogenetic relations between species. Graph-based methods which are suitable for two or more complete genomes rely on pairwise sequence similarities calculated between all sequences involved and an operational definition of orthology, for instance, reciprocal best hits. Some graph-based methods use clustering technique (for example single-linkage, complete-linkage, Markov Clustering algorithm) in order to extend nearest neighbor to more than two species and construct multi-species orthologous groups of particular granularity. Hybrid methods make use of both tree and graph representation at various state of processing, for example, to refine orthologous groups within a hierarchical framework of phylogenetic tree or guide to clustering procedure using a species tree (Chen, et al., 2007).

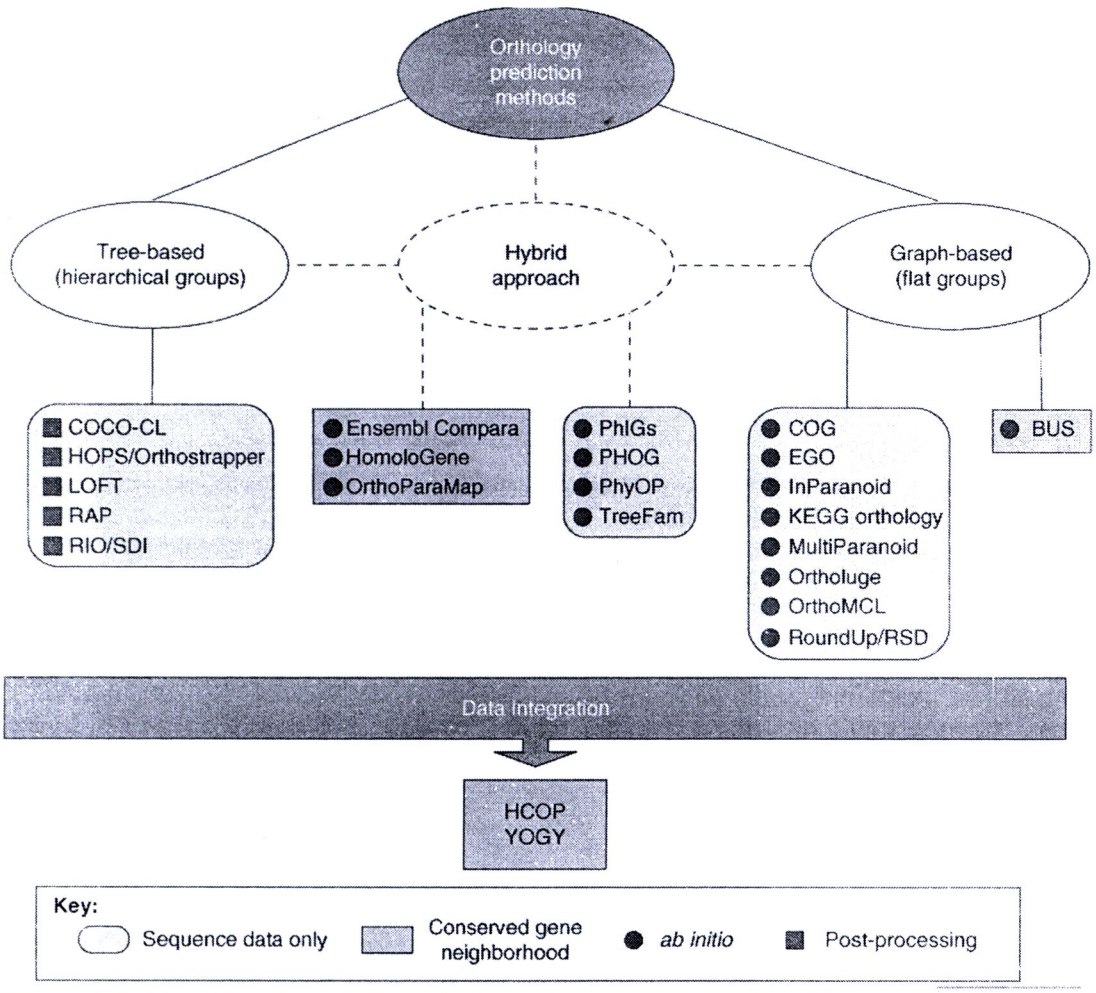

Figure 2.1 Classification of the orthology detection strategies. Three main categories are recognized according to the data representations they operate on, including tree-based, graph-based, and hybrid methods (Kuzniar, *et al.*, 2008).

Table 2.1 Comparison between various orthology and homology detection methods (Chen, *et al.*, 2007).

Methods	Strategy[a]	Apply to proteins	Grouping capability	Parameters analyzed	% positive protein pairs	
					Total[b]	Sampling average[c]
RIO	Phylogeny	Pfam domains	NO	Orthology bootstrap cutoff	1.9	17.9
Orthostrapper	Phylogeny	Pfam domains	NO	Orthology bootstrap cutoff	5.7	39.9
RSD	Distance	YES	NO	BLAST P E-value cutoff, Divergence cutoff	2.8	28.8
RBH	BLASTP	YES	NO	BLAST P E-value cutoff	5.2	37.7
Inparanoid	BLASTP	YES	YES (2 species)	BLAST P E-value cutoff	9.0	43.6
OrthoMCL	BLASTP	YES	YES	BLAST P E-value cutoff, MCL inflation index	11.8	56.6
KOG	BLASTP	YES	YES	N/A	23.6	66.2
SBH	Homology	YES	NO	BLAST P E-value cutoff	18.8	56.6
BLASTP	Homology	YES	NO	BLAST P E-value cutoff	41.5	72.1
TribeMCL	Homology	YES	YES	BLAST P E-value cutoff, MCL inflation index	47.2	74.7

[a]Alterrative orthology detection strategies (including phylogeny, distance or BLASTP-based), or homology detection methods.

[b]The fraction of positively predicted protein pairs (using default parameter settings) within the entire sampling dataset of 567,255 cross species homologous protein pairs (defined by Pfam domains).

[c]The average fraction of positively predicted protein pairs (using default parameter settings) from 100 sampling replicates (of the average total 1590 pairs)

Many strategies have been employed to distinguish the probable orthologs from paralogs, as summarized in the table 2.2. The phylogeny-based methods are including RIO (Resampled Inference of Orthology) and Orthostrapper/ Hierarchical grouping of Orthologous and Paralogous Sequences) (Storm and Sonnhammer, 2002). The methods which are based on evolutionary distance metrics are including RSD (Reciprocal Smallest Distance) (Wall, *et al.*, 2003). The BLAST-based methods are including the Reciprocal Best Hit (RBH), COG (Cluster of Orthologous Groups) (Tatusov, *et al.*, 2000; Tatusov, *et al.*, 2001) /KOG (euKaryotic Orthologous Groups) (Tatusov, *et al.*, 2003), and Inparanoid (Remm, *et al.*, 2001). An orthoMCL algorithm improves reciprocal best hit by the following reasons: (*i*) recognizing co-ortholog relationships, (*ii*) using a normalization step to correct for systematic biases, and (*iii*) using a Markov graph clustering (MCL) algorithm to define ortholog groups of proteins. The orthoMCL and Inparanoid exhibit a similar performance when comparing between two species, nevertheless, the former is extensible to cluster orthologs across multiple species. Analysis of independently could assign the enzyme categories (EC) number annotations which suggests a high degree of reliability, and orthology predictions for 55 genomes are available at the orthoMCL databases (OrthoMCL-DB) (Chen, *et al.*, 2006).

The COG/KOGs clusters are building in such a way that they are often contaminated with an out-paralogs. The InParanoid algorithm was specifically designed to find all in-paralogs in ortholog groups between two species. The MultiParanoid algorithm can assemble InParanoid clusters into multi-species groups. OrthoMCL was built in a fashion similar to InParanoid in terms of gathering inparalogs. A major difference however is that OrthoMCL provides the possibility of building ortholog groups of multiple species. Clustering of the orthologs is done by using the Markov Clustering algorithm (MCL), which is based on probability and graph flow theory. The ortholog assignments in the KEGG (Kyoto Encyclopedia of Genes and Genomes) database have a focus on similarity in molecular function. This is based on protein sequences comparison, information from the COGs database, and expert classifications of protein families. PhiGs used a graph-based method guided by known phylogenetic relationships to cluster orthologs. The orthologous clusters available in MGD at the research community Mouse Genome Informatics (MGI) were identified using a combination of computational analysis and manual curation. Most of the orthology assignments were extracted from the scientific publication. TreeFam is a manually curated database of trees with genes from animal taxa. The families were based on seed clusters from the PhIGs database that were expanded by both BLAST and HMM searching. A plant specific database called OrthologID was recently built from tree finishing plant genomes (Chiu, *et al.*, 2006). The gene family clusters were built from a BLAST hits and subjected to parsimony tree analysis. The HOPS database uses gene trees to extract orthologs from two species with the Orthostrapper algorithm. This method looks for ortholog groups between two species that cluster below an out-group, as illustrated in Table 2.3 (Deluca, *et al.*, 2006).

Table 2.2 Comparison of ortholog databases (Deluca, *et al.*, 2006).

Ortholog resource	COG/KOG	InParanoid/ MultiParanoid	OrthoMCL	HOPS
Number of species	66/7	22/4	55	n/a
Pros	Has become a standard for 'uniform-function' protein groups. Easy addition of new genome without recalculate the whole set. Manually curation. Provide species-specific expansion	Include genome for all major eukaryotic clades. Precise and exhaustive ortholog delineation for pairwise proteome analysis	Multiple species comparisons.	Doman oriented, integrated in the Pfam server Graphical user interface. Include also partially sequenced species.
Cons	Contain many outparalogs	No tree view provided. Only one Prokaryote (*E. coli*).	Some clusters contain outparalogs. Include multiple splice varients of genes	Only pairwise orthology between 2x3 eukaryotic clades. No prokaryote. Not downloadable, only runs in the web browser. Not queryable

Ortholog resource	KEGG	PhIGs	MGD	TreeFam
Number of species	355	34	21 (focus 3)	20 complete
Pros	Manually curated, taking into account function information for ESTs and incomplete genomes. Pathway links clusters.	Tree view provided.	Orthology classification supported by scientific publication. Expert knowledge considered.	Manually curated based on trees. Ortholog pairs can be downloaded. Use both known and novel data from Ensembl databases.
Cons	Generate unexpectedly large cluster	Poor website. Clusters not downloadable.	Linited in species. Mainly mouse, rat and human Varying quality.	Only contains animal taxa, with expression of some plant and fungal species used as outgroups.

2.1.2 Evolutionary analysis

The comparative genomics are normally used to uncover the evolution of bacterial genomes (Koonin, *et al.*, 2000) as well as an eukaryote gemones. The elementary events of gene evolution can be roughly classified in order as follows; (*i*) vertical descent (speciation) with modification; (*ii*) gene duplication, also followed by descent with modification; (*iii*) gene loss; (*iv*) horizontal gene transfer (HGT); and (*v*) fusion, fission, and other rearrangement of genes. Vertical descent and duplication might be considered as primary events of genome evolution and have been well recognized in the pre-genomic era. In contrast, the crucial evolution importance of gene loss, HGT, and gene rearrangement are among the major, fundamental generalizations of the emerging evolutionary genomics (Wolf, *et al.*, 2001). Therefore, the major goal of comparative genomic in the phylogenomic study is to reveal those evolutionary scenarios by using the available genomic sequences.

An evolutionary tree can be reconstructed in several ways, for example, using presence-absence of genomes in clusters of orthologous gene, conserving of local gene order (gene pairs) among prokaryote genomes, using parameter of identities distribution of probable ortholog, comparing of trees constructed from multiple protein families, analyzing of concatenated alignments of ribosomal proteins (Mirkin, *et al.*, 2003). The different methods to determine the phylogenetic tree have a different interpretation. The phylogenetic tree, which are based on presence-absence of genomes in orthologous clusters and the trees based on conserved gene pairs appear to be strongly affected by gene loss and horizontal gene transfer. So, this method represents the evolutionary of the whole genome. On the other hand, the phylogenetic tree, which has been reconstructed by using an individual protein, is more likely represent the evolutionary of that protein. This methods, therefore, is a protein specific evolution. However, if the highly conserved protein sequences such as ribosomal proteins are used, the phylogenetic tree which was reconstructed by this method can represent to the evolution of those organisms. The trees based on identity distributions for orthologs and particularly the tree made of concatenated ribosomal protein sequences seemed to carry a stronger and more robust phylogenetic signal.

Several studies use COGs and evolutionary scenario for assigning the gene gain and loss events in small group of genome. Marakova, *et al.* (2006) investigated nine lactic acid bacteria (LAB) genomes by using phylogenetic analysis, comparison of gene content across the group and reconstruct the ancestral gene sets indicated a combination of extensive gene loss and key gene acquisitions via horizontal gene transfer during the evolution of lactic acid bacteria with their habitats. The phylogenetic analysis was based on the concatenated alignment of ribosomal proteins. This result phylogenetic tree showed the improvement of resolution and robustness in order to explain the evolution this organismal group. The comparison of gene content across the of lactic acid bacteria by comparing the protein encoding genes in 12 sequenced *Lactobaillales* genome, available at the time of this analysis with the cluster of orthologous gene to create *Lactobaillales* – specific cluster of orthologous genes (LaCOG) the result in the close species (LaCOG) is more finer than the COG that analyses in all of species. Reconstruction of the gene loss and gene gain in evolution movement of lactic acid bacteria by using the weight parsimony algorithm, which have been proposed by Mirkin B.G., *et al.*(2003), suggested the predicted genome size of the lactic acid bacteria ancestor and the gained and lost gene related to their habitat and environment besides the horizontal gene transfer from neighbor living organism around

them. However, many biological events of gene loss and gene gain are required more discussions and more evidences to illustrate the real incident that occurred in evolutionary time (Makarova, *et al.*, 2006).

Marakova, *et al.* (2008) described the analysis of gene present in most of the currently available archaeal genome sequences in view of their classification in cluster of orthologous gene specific to the archaea (archaeal cluster of orthologous groups of proteins; arCOG). It represented an updated extension of previous comparative genomic analyses of COGs though exclusively devoted to the archaea. Consequently, the arCOG database produced is more refined, resulting in an increased coverage and resolution. The numerous growth of specific archaeal COGs and the accompanying decrease in the number of clusters containing paralogs were revealed. Thus, the comparison of the defined arCOGs allows to infer the presence of ~166 core arCOGs, which were likely present in the last archaeal common ancestor (LACA), while 282 and 336 arCOGs appear ancestral to the euryarchaeotal and crenarchaeotal branches, respectively. From the nature of the core arCOGs, the authors conclude that the LACA was a rather complex hyperthermophilic chemoautotroph possessing ~1000 genes. Differential gene gain and loss are predicted to have occurred in the two major archaeal branches. The pattern of arCOG distribution in the different archaeal genomes is used to reconstruct a gene-content tree. Although the type of analyses carried out is not innovative, the new arCOG database presented here will certainly be very useful to improve future genome annotations.

2.2 Cyanobacteria

Cyanobacteria are oxygenic photosynthetic bacteria that are widely distributed in aquatic and terrestrial environments, including extreme habitats such as hot springs, deserts, and polar regions. All cyanobacteria combine the ability to perform an oxygenic photosynthesis with typical prokaryote features (Ting, *et al.*, 2002). Nevertheless, some cyanobacteria can perform anoxygenic photosynthesis by using hydrogen sulfide (H_2S) as the electron donor (Cohen, *et al.*, 1986). Many cyanobacteria can fix atmospheric nitrogen (N_2) into a form of ammonia (NH_3), which the nitrogen is available for further biological reactions (Stal, 2009). Although a rather uniform in nutritional and metabolic respects, cyanobacteria are a morphologically diverse group of bacteria with a unicellular, filamentous, and colonial form. Unicellular and filamentous cyanobacteria can form symbioses with a wide diversity of hosts. In symbiosis, some cyanobionts perform both photosynthesis and nitrogen fixation while others exhibit only one of these properties (Zauner, *et al.*, 2006; Deusch, *et al.*, 2008; Ran, *et al.*, 2010). Moreover, it is generally accepted that the plastid in plants and algae are derived from a cyanobacterial ancestor after a long period of endosymbiosis, implicating the cyanobacteria in eukaryotic evolution. Without doubt, then, cyanobacteria are important to any understanding of Earth's early biological and environmental history (Tomitani, *et al.*, 2006).

Traditionally, the classification of the cyanobacteria is done by using the distinct morphology to divide the cyanobacteria group into five subsections according to table 2.3. Since there are vast amount of cyanobacterial genomes available, the classification of this organism have been changed to use the molecular marker instead of their morphology. For example, the use of 16S ribosomal RNA could classify cyanobacteria into seven clades (Honda, *et al.*, 1999). However, some studies argued that this is not sufficient for the study at a sub-generic level. Then, other genetic makers such as other cyanobacterial proteins, or the photosynthetic proteins were introduced for determining the phylogenetic tree at a sub-generic level have been purposed (Han, *et al.*, 2009). Furthermore, the availability of complete genome sequences in each group of cyanobacteria provides the opportunity to reconstruct biological events of genomic evolution through the analysis of functional core genes for example ribosomal protein genes (information processing protein) (Shi and Falkowski, 2008).

After the first eight cyanobacterial genomes had been completely sequenced, the ability to examine the distribution of genes in a very detailed way such as the finding of the signature genes in the cyanobacterial groups was proposed (Martin, *et al.*, 2003). From that study, hundreds of genes that are shared by eight cyanobacterial genomes were presented as core genes, which are a necessary gene for an entire group of cyanobacteria. Of those core genes, 181 genes have not been found to have any obvious homolog or ortholog in non-cyanobacterial bacterial genomes, whether photosynthetic or not. Therefore, those 181 genes are a signature core gene for cyanobacteria. These genes likely accounts for the unique shared characteristics of the cyanobacterial phenotype and are therefore a characteristic of the cyanobacteria groups. By using the same aspect, the signature gene can be discovered for each specific clade. There is a study that showed the cyanobacterial signature genes for each specific clade by using the comparative genome analysis of 44 cyanobacterial genome and revealed that there 39 proteins that are specific for all cyanobacteria (Gupta and Mathews, 2010).

Shi and Falkowski (2008) attempted to identify the stable core gene from those core genes, which have been found in all cyanobacterial genome. The highly conserved genes involve in the important mechanisms and components of cyanobacteria, especially the photosynthetic and ribosomal apparatus. Moreover, the stable core gene was also hypothesized that are more resistant to the horizontal gene transfer (HGT) than other genes. In the divergent manner, the variable shell genes are gene that more susceptible to the horizontal gene transfer. Identification of core gene potentially allows separation of true phylogenetic signal from "noise". Shi and Falkowski (2008) also demonstrated the overall of phylogenetic incongruence among 682 orthologous protein families from 13 genomes of cyanobacteria. The principle coordinates analysis (PCoA) was used to discover the core gene set consisting of 323 genes with similar evolution paths. The reconstructed phylogeny of 13 genomes from concatenated 323 core proteins by using three methods resulted in the same topologies as that for the consensus, supertree and concatenated of all 682 proteins families but it can show the high resolution of the evolutionary event (Shi and Falkowski, 2008).

Table 2.3 Characteristics of the Cyanobacteria Subsections by using the morphological approaches (Garrity, *et al.*, 2005).

Sub Section	General Shape	Reproduction and Growth	Heterocyst	% G+C	Other Properties	Representative Genera
I	Unicellular rods or cocci; nonfilamentous aggregates	Binary fission, budding	-	31 – 71	Almost always nonmobile	Chamaesiphon Chroococcus Gloeothece Gleocapsa Prochloron
II	Unicellular rods cocci; may be held together in aggregates	Multiple fission to from baeocytes	-	40 – 46	Only some baeocytes are mobile	Pleurocapsa Dermacarpa Chroococciopsis
III	Filamentous, unbranched trichome with only vegetative cell	Binary fission in a single plane fragmentation	-	34 – 67	Usually motile	Lyngbya Oscillatoria Prochlorothrix Spirulina Pseudanabaena
IV	Filamentous, unbranched trichome may contain specialized cells	Binary fission in a single plane, fragmentation to form homogonia	+	38 – 47	Often motile, may produce akinetes	Anabaena Cylindrospernum Aphanizomenon Nostoc Scytonema Calothrix
V	Filamentous trichomes or compose of more than one row of cell	Binary fission in more than plane, homogonia form	+	42 – 44	May produce akinete, greates morphological complexity and differentiation in cyanobacteria	Fischerella Stigonema Geitleria

2.2.1 Photosynthesis in cyanobacteria

Photosynthesis is the most important biological process on Earth, to convert the solar energy to the chemical energy in the photosynthetic organisms. The input is carbon dioxide (CO_2), water (H_2O), minerals and light, and the outputs are carbohydrates, and oxygen (O_2). This oxygenic photosynthesis occurs in higher eukaryotes such as in plants, as well as in a photosynthetic prokaryotes (Buick, 2008). The photosynthetic apparatus are distributed in six prokaryotic phyla i.e. *Chlorobi*, *Chloroflexi*, *Heliobacteria*, *Proteobacteria*, *Firmicutes* and cyanobacteria (Bryant and Frigaard, 2006). Nevertheless, the question about the evolution phylogenetic apparatus is still ambiguous. Recent study used the geological and molecular phylogenetic evidences to suggest several alternative evolutionary scenarios of the origin of photosynthesis in those photosynthetic organisms (Xiong, 2006). However, the photosynthetic apparatus in the prokaryotic groups also different, such as the reaction center types and antenna systems. Thus, the photosynthetic pathway, also the photosynthetic reactions for each photoautotroph organism are wildly different. The reference photosynthetic pathway and the antenna protein (light harvesting complex) have been illustrated in the figures 2.1 and 2.2, respectively.

For the general photosynthesis in cyanobacteria, phycobilisome proteins serve as the primary light-harvesting antennae for the Photosystem II, whereas some marine picocyanobacteria, such as *Prochlorococcus* use a chlorophyll a_2/b_2 light-harvesting complex. Ting, *et al.* (2002) presented a scenario to explain how the *Prochlorococcus* antenna might evolve in an ancestral cyanobacterium in iron-limited oceans, resulting in the diversification of the *Prochlorococcus* and marine *Synechococcus* lineages from a common phycobilisome-containing ancestor. Differences in the absorption properties and cellular costs between chlorophyll a_2/b_2 and phycobilisome antennas in extant *Prochlorococcus* and *Synechococcus* appear to play a role in differentiating their ecological niches in the ocean environment. Here, the genomic information, which have been carried along generation to generation, can suggest the scenarios that happened during the evolution of these organism.

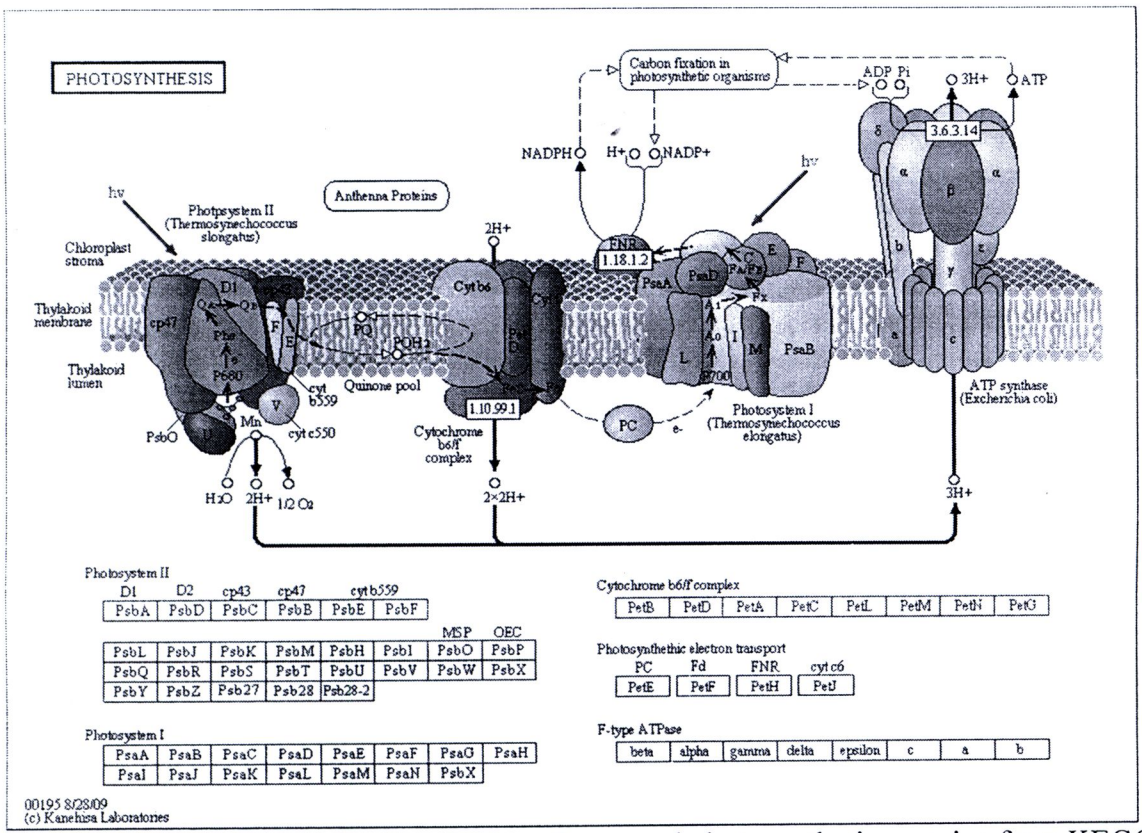

Figure 2.2 The reference photosynthesis pathway and photosynthetic proteins from KEGG databases (http://www.genome.ad.jp/kegg/).

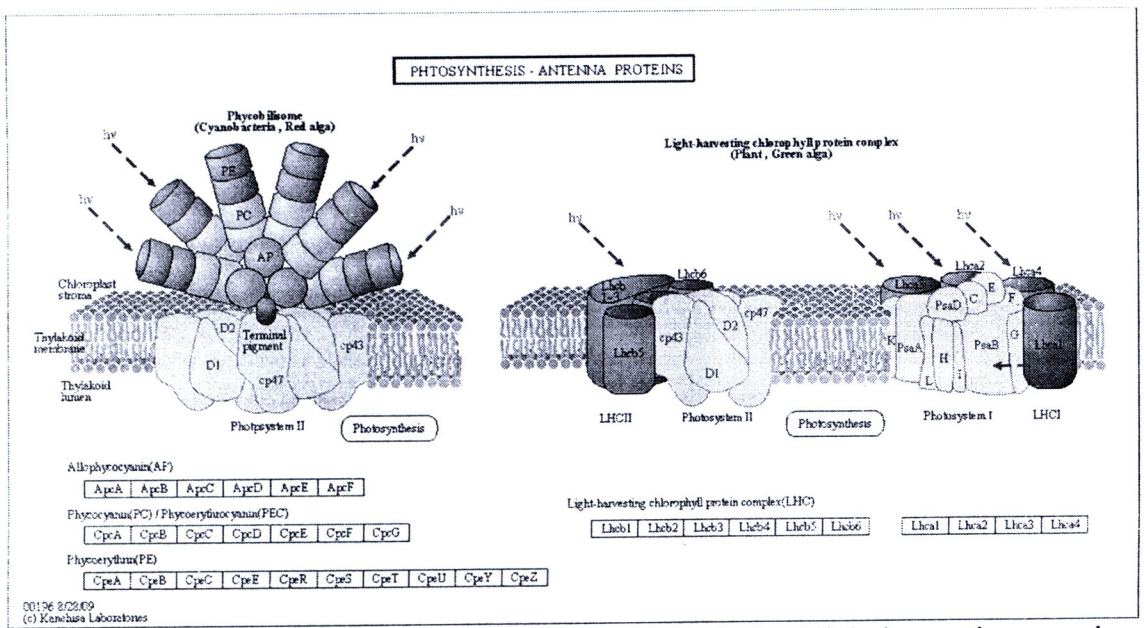

Figure 2.3 The reference photosynthesis antenna proteins and light harvesting complex from KEGG databases (http://www.genome.ad.jp/kegg/).

2.2.2 Evolutionary study in cyanobacterial lineages

In order to gain better understanding of cyanobacterial genomes and their associated genomic functions, the comparative genomic analysis and evolutionary study have been introduced to uncover this linkage. Using the enormous genomic information, the evolutionary study can be used to identify the diversity and complexity of the interested genomes. Then, the core genes, which are resistant to the horizontal gene transfers, and shell genes, which are susceptible to the horizontal gene transfers, of cyanobacteria could be determined by using the phylogenomic study (Shi and Falkowski, 2008). Additionally, the specific genes, which mean the genes that necessary for some cyanobacterial lineages for example the genes that specific for habit in fresh water, ocean, low light intensity, high light intensity, or in the extreme environment, could be explained by using evolutionary study (Gupta and Mathews, 2010). Therefore, the advancement of phylogenetic analysis can be used for determine the evolutionary lineages, the specific core genes, and the signature gene for some cyanobacterial clades.

Typically, the ecological and environmental niches are the major driving force of the evolutionary divergences (Kunin and Ouzounis, 2003). For instance, organisms, which are living in the deep ocean and coastal, are usually needed a different essential genes. Those genes depend on the difference environmental factors such as light, nutrient, salinity, or even the neighbor organisms. All of environmental factors are a driven pressure for the organismal evolution. Then, it is obvious that the diverse genotypes are driven from the environmental niches. In cyanobacterial evolution, several studies have been done in the small cyanobacterial groups with the difference approaches in order to explain the driving force of the environmental niches. For example, a recent study attempted to explain the lifestyle of marine cyanobacteria by using the phylogenomic study (Dufresne, *et al.*, 2008). The rich-nutrient and poor-nutrient are a main driving pressure for the difference genomic contents. Furthermore, the evolutionary study could be used for tracing the origin of photosynthesis in photosynthetic bacteria (Mulkidjanian, *et al.*, 2006). In conclusion, it have been proofed that a main reason for the divergence and complexity of organism are driven from the environmental pressure.