January 2005

TO: U.S. Department of Energy
Office of Science

RE: **Final Report**

PROJECT TITLE: An integrative approach to energy, carbon, and redox
metabolism in the cyanobacterium *Synechocystis* sp. PCC 6803

PRINCIPAL INVESTIGATOR: Ross Overbeek
(630)-567-7677

RECIPIENT
ORGANIZATION: Fellowship for Interpretation of Genomes
15W 155 81st Street
Burr Ridge, Illinois 60527

DOE AWARD NUMBER: DE-FG02-04ER15499

PROJECT PERIOD: from 11/15/2003 through 11/14/2004

BUDGET PERIOD
COVERED BY THIS REPORT: from 11/15/2003 through 11/14/2004

Signature of Principal Investigator

NOTE:

This project is a part of a collaborative research effort involving three other groups lead by:
Willem F.J. Vermaas (Arizona State University) coordinating PI
Julian Whitelegge and Kym Faull (University of California, Los Angeles)
Robert W. Roberson (Arizona State University)

Final Report

Overview

The team of the Fellowship for Interpretation of Genomes (FIG) under the leadership of Ross Overbeek, began working on this Project in November 2003. During the previous year, the Project was performed at Integrated Genomics Inc. A transition from the industrial environment to the public domain prompted us to adjust some aspects of the Project. Notwithstanding the challenges, we believe that these adjustments had a strong positive impact on our deliverables. Most importantly, the work of the research team led by R. Overbeek resulted in the deployment of a new open source genomic platform, the SEED (**Specific Aim 1**). This platform provided a foundation for the development of CyanoSEED a specialized portal to comparative analysis and metabolic reconstruction of all available cyanobacterial genomes (**Specific Aim 3**). The SEED represents a new generation of software for genome analysis. Briefly, it is a portable and extendable system, containing one of the largest and permanently growing collections of complete and partial genomes. The complete system with annotations and tools is freely available via browsing or via installation on a user's Mac or Linux computer. One of the important unique features of the SEED is the support of metabolic reconstruction and comparative genome analysis via encoding and projection of functional subsystems. During the project period, the FIG research team has validated the new software by developing a significant number of core subsystems, covering many aspects of central metabolism (**Specific Aim 2**), as well as metabolic areas specific for cyanobacteria and other photoautotrophic organisms (**Specific Aim 3**). In addition to providing a proof of technology and a starting point for further community-based efforts, these subsystems represent a valuable asset. An extensive coverage of central metabolism provides the bulk of information required for metabolic modeling in *Synechocystis* sp. PCC 6803. Detailed analysis of several subsystems covering energy, carbon, and redox metabolism in the *Synechocystis* sp. PCC 6803 and other cyanobacteria has been performed (**Specific Aim 4**).

The main objectives for this year (adjusted to reflect a new, public domain, setting of the Project research team) were:

Aim 1. To develop, test, and deploy a new open source system, the SEED, for integrating community-based annotation, and comparative analysis of all publicly available microbial genomes. Develop a comprehensive genomic database by integrating within SEED all publicly available complete and nearly complete genome sequences with special emphasis on genomes of cyanobacteria, phototrophic eukaryotes, and anoxygenic phototrophic bacteria - invaluable for comparative genomic studies of energy and carbon metabolism in *Synechocystis* sp. PCC 6803.

Aim 2. To develop the SEED's biological content in the form of a collection of encoded Subsystems largely covering the conserved cellular machinery in prokaryotes (and central metabolic machinery in eukaryotes).

Aim 3. To develop, utilizing core SEED technology, the CyanoSEED - a specialized WEB portal for community-based annotation, and comparative analysis of all publicly available cyanobacterial genomes. Encode the set of additional subsystems representing key metabolic transformations in cyanobacteria and other photoautotrophs. We envisioned this resource as complementary to other

public access databases for comparative genomic analysis currently available to the cyanobacterial research community.

Aim 4. Perform in-depth analysis of several subsystems covering energy, carbon, and redox metabolism in the *Synechocystis* sp. PCC 6803 and all other cyanobacteria with available genome sequences. Reveal inconsistencies and gaps in the current knowledge of these subsystems. Use functional and genome context analysis tools in CyanoSEED to predict, whenever possible, candidate genes for inferred functional roles. To disseminate freely these conjectures and predictions by publishing them on CyanoSEED (<http://cyanoseed.thefig.info/>) and the Subsystems Forum (<http://brucella.uchicago.edu/SubsystemForum/>) in order to facilitate experimental analysis by our collaborator on this Project and by other experimentalists working in various field of cyanobacterial physiology and biotechnology.

Specific Aim 1

The FIG in collaboration with Argonne National Lab has developed and launched initial versions of the new open source genomic integration platform, the SEED, funded partially by this grant. The SEED represents a new generation of public web-based environments supporting community-based genome annotation and subsystems analysis. Several research teams (including groups in Bielefeld University, Germany, UCSD and the Burnham Institute, CA) are currently contributing to testing and improvement of the system. The SEED resource has been installed at ~30 locations and now forms a framework, which can be accessed for example at: <http://theseed.uchicago.edu/FIG/index.cgi> (**Figure 1**). Protocols for synchronization of autonomous installations of the SEED have been developed supporting peer-to-peer (p2p) updates of code, raw data and annotations. The SEED supports:

- semi-automated genome comparative analysis and annotation,
- pathway and subsystem reconstruction and analysis across multiple species,
- community annotations and alternative assignments from major public integrations;
- gene discovery via the use of genome context analysis techniques,
- integration, comparative analysis and interpretation of functional genomics data.

The current integration contains data from 470 bacterial (of these 261 are complete or near completion), 32 archaeal (21 more or less complete), 558 (16 complete) eukaryotic, and 1272 viral genomes, including complete and nearly complete genomes of 14 cyanobacteria. The major effort at FIG is dedicated to continually adding new genomes to this collection, keeping pace with the avalanche of newly available sequence data.

Specific aim 2.

In addition to and in parallel with the collaborative effort on infrastructure and tool development, during the past year the FIG team has been developing SEED biological content: curating annotations, encoding the core metabolism across all available genomes, identifying “missing genes” and other open problems in our current understanding of cellular processes, and generating functional conjectures. This work in SEED is uniquely built on the **subsystem approach** to metabolic reconstruction and comparative genome analysis. The term “Subsystem” in SEED refers to a set of functional roles (e.g. enzymes) that together implement a specific cellular function (metabolic or non-metabolic) or a structural complex across large sets of genomes. It may be thought of as a generalization of a pathway concept. By the term *annotated subsystem* we refer to a spreadsheet, in which each column represents a functional role for the subsystem, each row

represents a specific genome, and each cell contains those genes from the specific organism that implement the specific functional role. While defining a subsystem, annotators include a collection of functional roles broad enough to cover distinct **variants** in all relevant organisms, since organisms may harbor genes that implement slightly different subsets of roles that make up the same subsystem. Each subset of functional roles that exists in at least one organism with an operational version of the subsystem constitutes an operational variant.

We believe that a key to improved accuracy in genome annotation and analysis, is to have **experts annotate single subsystems over the complete collection of genomes**. This approach differs sharply from the common annotation methodology, when teams analyze a single genome at a time, ensuring that annotators have no special expertise relating to the vast majority of genes they annotate. By having individuals annotate single subsystems over a large collection of genomes, SEED allows curators with expertise in specific pathways (or, more generally, subsystems) to perform their task with relatively few errors and high efficiency. Furthermore, advantages of the subsystem-based strategy extend beyond initial genome annotations. In-depth analysis of individual subsystems allows researchers to get significant new insights into specific cellular processes across multiple genomes including: (i) refinement of tentative annotations, (ii) exposure of numerous errors and open questions in current understanding of a subsystem; (iii) identification of “missing” genes, (iv) cataloging subsystem variations in different species. Such analysis provides an effective guidance to planning further experimental efforts to clarify issues relating to a subsystem, as well as underlying data needed to support a detailed review of a subsystem. **Appendix 1** describes in detail the process of initial encoding and in-depth analysis of a subsystem using NAD and NADP cofactor biosynthesis in cyanobacteria as an example.

During the funded period, over 170 subsystems that cover most of the conserved cellular machinery in prokaryotes (and central metabolic machinery in eukaryotes) have been initially encoded and largely populated with the corresponding sets of genes in 315 prokaryotic and eukaryotic genomes. Together they include over 1200 non-redundant functional roles. In the context of this Project, a special emphasis was made on curation of *Synechocystis* sp. PCC 6803 and other photoautotrophic genomes. The quality of the emerging body of annotated subsystems varies greatly. A limited set have been done by domain experts and reflect a substantial research background. Many of the subsystems have been done as drafts in attempts to gain accurate overviews of specific topics; often these produced key insights, but are in need of refinement and extension. An up-to-date list of available subsystems (with detailed evaluation of the quality and the stage of completion of each) is maintained at the Subsystems Forum (<http://subsys.info>).

Specific aim 3.

The SEED technology became the foundation for the development of **CyanoSEED** - a specialized WEB portal for community-based annotation and comparative analysis of all publicly available cyanobacterial genomes (**Figure 2**). CyanoSEED (<http://cyanoseed.thefig.info/>) features access to all internal SEED tools and collection of Subsystems, as well as links to high-impact public genomic recourses (including UniProt, KEGG, COG, RefSeq, TIGR) and to specialized cyanobacterial sites: CyanoBase, CyanoSite, and CyanoNews. CyanoSEED integrates (among hundreds of bacterial, archaeal, eukaryotic, and viral genomes) all currently available complete and nearly complete genomes of cyanobacteria, as well as anoxygenic phototrophic bacteria and higher

plants – important for comparative genomic studies of energy and carbon metabolism in *Synechocystis* sp. PCC 6803. In addition to the core metabolic subsystems in SEED, many new subsystems representing key cellular processes characteristic of cyanobacteria and higher plants have been added to CyanoSEED, including composition of photosynthetic and respiratory membrane complexes, inorganic carbon concentration and fixation mechanisms, vital pigment and cofactor biosynthetic pathways (phycobilins, carotenoids, chlorophylls, quinones), plant hormones and secondary metabolites, circadian clock, cyanophycin metabolism, and many others. Out of the 3566 ORFs identified in the *Synechocystis* 6803 genome, 1843 (52%) have been assigned with non-hypothetical functions in SEED, and 720 have been included in subsystems so far -- with curation work ongoing. **Appendix II** contains several examples of such subsystems and exemplifies their application for detailed cross-genome analysis of a pathway, for detection of unresolved questions, as well as for generating conjectures that can be tested experimentally. The subsystem collection arranged by areas of metabolism can be accessed at <http://cyanoseed.thefig.info/>. Subsystem names are hyperlinked to their actual encodings in SEED. A globe icon near a Subsystem name indicates a link to a corresponding Subsystem diagram (if available).

Cyanobacteria-focused subsystems together with the core metabolic subsystems in the SEED constitute what we call the informal reconstruction of the *Synechocystis* 6803 metabolism. It currently includes the majority of the components of a draft reconstruction generated in the year 2003 (within the ERGO platform) – 720 protein-encoding genes. Development of a new reconstruction that will provide sufficient coverage of *Synechocystis* 6803 metabolism to initiate stoichiometric modeling effort has begun in collaboration with the lab of Bernard Palsson (University of California, San Diego) – the leading group in the field of metabolic modeling.

Although this collection of subsystems is of substantial value, we believe that the main impact of CyanoSEED is in providing the cyanobacterial community with a framework for capturing and analysis of a rapidly growing body of genomic and postgenomic data. We anticipate, that researchers will develop and share their own collections of annotated subsystems in the specific areas of their research. We realize that, besides SEED, several integrated databases containing phylogenetic occurrence, clustering or protein fusion data, and many combinations of the three have been implemented in the last five years (**Table 1**). However, several important features distinguish the CyanoSEED from other critical resources (such as KEGG or CyanoBase) available for cyanobacterial research community. First, SEED software supports metabolic reconstruction and comparative genome analysis via encoding of Subsystems (as first class objects). The utility of this unique powerful approach to genome annotation and comparative analysis has been briefly described above and is illustrated in **Appendix I** and **II**. Another important SEED feature is that it is readily editable and expandable by an inexperienced user. Editing of existing subsystems and construction of new ones does not require any programming skills, providing experimental biologists with unique opportunities of fully interactive *in silico* analysis of metabolic pathways at the whole-genome scale. It is our believe, that CyanoSEED can be easily adopted by researchers working on various aspects of cyanobacterial biology and biotechnology, allowing them to effectively address research topics related to their subsystems of interest. We anticipate that due to the unique combination of subsystem-based metabolic reconstruction technology with various techniques of genome context analysis, supported by the SEED platform, this work will become one of the key common denominators for research scientists wishing to extract value from genomes, inspiring a completely new way of reasoning about functions of genes and gene ensembles.

Figure 2. The CyanoSEED main page (fragment).

Table 1: Public access databases for comparative genomic analysis

Organism specific	Reference
<i>CyanoBase</i>	http://www.kazusa.or.jp/cyano/
<i>Arabidopsis thaliana</i>	http://www.arabidopsis.org/
<i>Escherichia coli Ecogene</i>	http://bmb.med.miami.edu/EcoGene/EcoWeb/
<i>Bacillus subtilis Micado</i>	http://genome.jouy.inra.fr/cgi-bin/micado/index.cgi
<i>S. cerevisiae</i>	http://www.yeastgenome.org/index.html
<i>C. elegans</i>	http://www.wormbase.org/
<i>Drosophila melanogaster</i>	http://flybase.bio.indiana.edu/
Comparative analysis	Reference
Cluster of Orthologous Groups (COG)	http://www.ncbi.nlm.nih.gov/COG/
FusionDB and PhydBac	http://igs-server.cnrs-mrs.fr/phydbac/ http://igs-server.cnrs-mrs.fr/FusionDB/
Microbial Database (MDB)	http://www.tigr.org/tdb/mdb/mdbcomplete.html
Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)	http://string.embl.de/
GenomeNet and KEGG	http://www.genome.ad.jp/
MetaCyc	http://metacyc.org/

Specific Aim 4.

Analysis of several subsystems covering energy, carbon, and redox metabolism (the focus of this Project) in *Synechocystis* 6803 and all other cyanobacteria with available genome sequences has been conducted in order to reveal inconsistencies and gaps in the current knowledge of these metabolic areas, and to generate experimentally testable hypothesis and conjectures deepening our understanding of these complex cellular processes. Please, see **Appendix II** for detailed discussion of these subsystems. This analysis was greatly facilitated by **genome context analysis tools** supported by SEED. Unlike the majority of annotation tools, genome context analysis techniques are not based on sequence similarity. Their utility in predicting gene functions grows dramatically with the number of genomes used in analysis. The ongoing explosion in available genome sequences has led to rapid growth of their significance in genome annotation and gene discovery efforts (for recent overviews see (Koonin 2002; Huynen, Snel et al. 2003) (Osterman and Overbeek 2003). Some context analysis techniques currently supported in SEED include:

1. Clustering on a chromosome – hugely important for inferring “functional coupling” between genes. It is based on the observation that functionally related genes tend to cluster on prokaryotic chromosomes ((Galperin and Koonin 2000). Precomputed cases of functional coupling of genes are indicated in SEED by blue asterisks (Figure 3 (1)). In addition, the “Get detailed Functional Coupling Data” tool can be evoked from every protein page to recalculate functional coupling on the fly (Figure 3, (2)).

The SEED: an Annotation/Analysis Tool Provided by [FIG](#)

[Subsystem Forum](#) | [FIG Tutorials](#) | [Peer-to-peer Updates](#) | [SEED Control Panel](#) | [SEED Wiki](#)
[GOLD](#) | [NCBI](#) | [KEGG](#) | [ExPASy](#) | [TIGR cmr](#) | [UniProt](#) | [Report Bug?](#)

SEED version cvs.1106500526 (Sun Jan 23 11:15:26 2005) on seed-linux-2.uchicago.edu

[FIG search](#)

Protein [fig|1148.1.peg.2724](#): *Synechocystis* sp. PCC 6803

[Translate Function Assignments](#)

Context on contig NC_000911

fid	starts	ends	size	gap	req. in pin	fc	neigh	comment	aliases	Related
2718	3040804	3041412	609	+				hypothetical protein	GeneID:952788, NP_442617.1 , gil16331889 , kegg syn:slr0318 , slr0318 , trjQ55925 , unilQ55925	
2719	3041494	3042405	912	+	81			beta-lactamase	GeneID:952781, NP_442618.1 , blaOXA-3 , gil16331890 , kegg syn:slr0319 , slr0319 , trjQ55926 , unilQ55926	
2720	3042684	3042490	195	-	84			unknown protein	GeneID:952222, NP_442619.1 , gil16331891 , kegg syn:ssl0511 , ssl0511 , trjP74803 , unilP74803	
2721	3042789	3043610	822	+	104			Methionine aminopeptidase (EC 3.4.11.18)	GeneID:952221, NP_442620.1 , gil16331892 , kegg syn:slr0786 , map , slr0786 , sp P53580 , unilP53580	
2722	3043926	3043531	396	-	80			unknown protein	GeneID:952220, NP_442621.1 , gil16331893 , kegg syn:slr0800 , slr0800 , trjQ55927 , unilQ55927	
2723	3044833	3045849	1017	+	906			Nicotinamide-nucleotide adenyltransferase (EC 2.7.7.1) / ADP-ribose pyrophosphatase (EC 3.6.1.13)	GeneID:952219, NP_442622.1 , gil16331894 , kegg syn:slr0787 , slr0787 , sp Q55928 , unilQ55928	
2724	3045877	3047262	1386	+	27			Nicotinamide phosphoribosyltransferase (EC 2.4.2.12)	GeneID:952218, NP_442623.1 , gil16331895 , kegg syn:slr0788 , slr0788 , trjQ55929 , unilQ55929

Subsystems in which this peg is present

Subsystem	Role
NAD_and_NADP_cofactor_biosynthesis_global	Nicotinamide phosphoribosyltransferase (EC 2.4.2.12)

[To View Annotations](#) / [To View All Related Annotations](#)
[Protein Sequence](#)
[DNA Sequence](#)
[To Make an Annotation](#)
[To Get Detailed Functional Coupling Data](#) 2
[To Get Fusion Data](#) 3
[To Find Homologs in Clusters](#)
[To Compare Region](#)

Similarities

MaxN: 50

Max expand: 5

Figure 3. A sample gene/protein page in SEED (fragments) displays genome environment, internal and external annotations, links to external databases, and provides access to similarity tables, genome context analysis and annotation tools.

2. Protein (domain) fusion events provide another form of evidence of potential functional coupling. In this technique, originally termed “Rosetta stone” (Marcotte, Pellegrini et al. 1999), one searches for a pair of genes in one organism that appear to be fused into a single gene in a second organism (Figure 3, (3)).
3. Occurrence (phyletic) profiling (Pellegrini, Marcotte et al. 1999) generates instances of potential functional coupling for a pair of proteins based on their simultaneous presence or absence in a subset of genomes in correlation with the presence or absence of a particular functional variant of a subsystem (Appendix II, Figure 4B).

One of the many interesting outcomes of in-depth Subsystems analysis is the prediction of “missing genes.” “Missing genes” are cases when the presence of a functional role in an organism or a group of organisms is inferred by Subsystem analysis, but a respective gene can’t be detected in the corresponding genome(s) by a reliable homology-based projection (reviewed in (Osterman and Overbeek 2003)). Filling in missing pieces in the energy, carbon, and redox metabolism of cyanobacteria – prominent players in the global carbon cycle – is a fundamentally important task. FIG team was able to suggest candidate genes for some of the inferred functional roles in *Synechocystis* 6803 and other experimentally tractable cyanobacteria (**Table 2**) based on the combination of advanced techniques of the genome context and functional context analysis supported by SEED. The type of analysis performed at this stage is illustrated in **Appendix II** (using Carotenoids biosynthesis as an example) and in **Appendix I** (NAD and NADP biosynthesis in cyanobacteria). The complete and ever-growing log of “missing genes” and predictions (wherever available) is maintained on the Subsystems Forum (<http://brucella.uchicago.edu/SubsystemForum/>), which can be accessed from CyanoSEED WEB portal. We anticipate that experimental verification of selected predictions will be carried out in the lab of Dr. Vermaas, the PI on this grant, as well as by broader cyanobacterial community.

Table 2: "Missing genes" in complete and nearly complete cyanobacterial genomes (preliminary analysis).

Subsystem	Functional role	"Missing" in genomes	Gene candidates	
			Example	Key evidence
Succinate Dehydrogenase*	Succinate dehydrogenase cytochrome b-556 subunit	all	SY-slr0201, PC-pmt0131	
Succinate Dehydrogenase*	Succinate dehydrogenase anchor subunit SdhD	all	search in progress	
Coenzyme A Biosynthesis	Aspartate 1-decarboxylase (EC 4.1.1.11)	SE, PC, TE, SCW	search in progress	putative decarboxylases, occurrence profiling
Chlorophyll Biosynthesis	Chlorophyll a, chlorophyllide a oxygenase (EC 1.14.13.-)	PC	PC-pmm0808	long-range similarity, putat. [2Fe-2S] binding domain
Chlorophyll Biosynthesis	4-vinyl-protochlorophyllide reductase, BchJ (EC 1.3.1.75)	all	search in progress	
Porphyrin, Heme, and Siroheme	Protoporphyrinogen IX oxidase (aerobic and anaerobic)	all but TE & GV	search in progress	
Lysine Biosynthesis DAP Pathway	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-acetyltransferase	all	SY-sll1636	long-range similarity
Lysine Biosynthesis DAP Pathway	N-acetyl-L,L-diaminopimelate aminotransferase (EC 2.6.1.-)	all	SY-slr1022	ArgD homolog
Lysine Biosynthesis DAP Pathway	N-acetyl-L,L-diaminopimelate deacetylase (EC 3.5.1.18)	all	SY-slr1653, or sll0100	homology (paralogs)
Histidine Biosynthesis	Histidinol-phosphatase (EC 3.1.3.15)	all	SY-sll0084, TE-tlr1657	clustering, putative phosphatases
Fatty Acid Biosynthesis FASII	Holo-[acyl-carrier protein] synthase (EC 2.7.8.7)	all	SY-slr0495	clusters, yhhU/sfp homologs
Carotenoid biosynthesis*	Lycopene beta cyclase (EC 1.14.-.-)	all but PC, SCW, GV	SY-slr0941, or sll0513, or slr1579, or sll0659, or slr1394	long-range similarity, occurrence profiling
Carotenoid biosynthesis	Beta-carotene ketolase (EC 1.14.-.-)	TE, CW, TER, SE	search in progress	
Carotenoid biosynthesis	Hydroxyneurosporene synthase	"globally missing"	search in progress	
Carotenoid biosynthesis	Hydroxyneurosporene methyltransferase (EC 2.1.1.-)	"globally missing"	search in progress	
Plastoquinone Biosynthesis*	4-hydroxyphenylpyruvate dioxygenase (EC 1.13.11.27)	PC, SCW, TE, SE	search in progress	
Plastoquinone Biosynthesis*	Homogentisate prenyltransferase	PC, SCW, TE, SE	search in progress	
Thiamine Biosynthesis	Phosphomethylpyrimidine kinase (EC 2.7.4.7)	PC, SCW, SY, TE	search in progress	
Thiamine Biosynthesis	Conserved thiamin-related protein TenA	PC, SCW, SY, TE	search in progress	
Biotin Biosynthesis	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	SY, GV, NP, N, AV, TER	search in progress	
Biotin Biosynthesis	Biotin synthesis protein bioC	TE, SY, GV, NP, N, AV, TER	search in progress	
Biotin Biosynthesis	Biotin synthesis protein BioH	PC, SE, SCW, CW	PC-PMT1489; SE-Selo240901	clusters, occurrence profiling
Methionine biosynthesis*	Lack all genes involved in the homocysteine synthesis	TE, SY, GV, N, CW, TER	search in progress	Occurrence profiling, co-regulation, Clustering
Ribonucleotide reduction	Ribonucleotide reductase transcriptional regulator NrdR	all	slr0780 in SY	Occurrence profiling, co-regulation, Clustering
Cobalamin (vitamin B12) biosynthesis	Cobalt transporter (Co is involved in B12)	all but TE	slr2135 in SY	Regulation by B12-specific regulatory RNA element

Organism abbreviations are as follows: **SY** – *Synechocystis* 6803, **SE** – *Synechococcus elongatus* PCC 7942, **PC** - all *Prochlorococcus marinus* genomes, **TE** – *Thermosynechococcus elongatus* BP-1, **TER** - *Trichodesmium erythraeum* IMS101 **SCW** - *Synechococcus* WH 8102, **GV** - *Gloeobacter violaceus* PCC 7421. CyanoBase gene IDs are used for inferred candidates. Cases labeled with a star (*) are discussed in detail in **Appendix II**.

References:

- Galperin, M. Y. and E. V. Koonin (2000). "Who's your neighbor? New computational approaches for functional genomics." Nat Biotechnol **18**(6): 609-13.
- Huynen, M. A., B. Snel, et al. (2003). "Function prediction and protein networks." Curr Opin Cell Biol **15**(2): 191-8.
- Koonin, E. V., Galperin, M.Y. (2002). SEQUENCE - EVOLUTION - FUNCTION. Computational Approaches in Comparative Genomics. Boston, Kluwer Academic Publishers.
- Marcotte, E. M., M. Pellegrini, et al. (1999). "Detecting protein function and protein-protein interactions from genome sequences." Science **285**(5428): 751-3.
- Osterman, A. and R. Overbeek (2003). "Missing genes in metabolic pathways: a comparative genomics approach." Curr Opin Chem Biol **7**(2): 238-51.
- Pellegrini, M., E. M. Marcotte, et al. (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." Proc Natl Acad Sci U S A **96**(8): 4285-8.

Appendix I

An example of subsystem-based analysis: NAD and NADP biosynthesis in *Synechocystis* sp. PCC 6803 and other cyanobacteria.

1. Encoding a subsystem.

Encoding of the subsystem in SEED is initiated by building a table of relevant functional roles, such as shown for the **NAD and NADP cofactor biosynthesis** subsystem by the rows of Figure 1,A. The roles are defined by the most standard descriptive names, for example enzyme names and corresponding Enzyme Classification (EC) numbers, whenever they are available. These names must exactly match the gene annotations in the underlying database. To do so, SEED annotation system provides variety of annotation tools (Figure 2), which allow the authenticated user to assign any names for the genes and seamlessly propagate them to the whole set of selected orthologs. Annotation tools are integrated with the tables of pre-computed similarities, external annotations (from Swiss-Prot, TREMBL, PIR, GenBank, and CyanoBase) and alignment/trees visualization tools. This allows user to examine and evaluate the evidence prior to propagating functional annotation.

Once the initial set of functional roles is defined and matched with annotations of at least one basic model organism (in our case, *Synechocystis* 6803), the system automatically fills in (Figure 2, (1)) the gene/protein unique identifiers (IDs) in the cells of a spreadsheet (Figure 4). Subsystems are dynamic, functional roles can be added, renamed or removed, as a result of adding more genomes (see below) or refining the underlying information.

The main encoding/editing module currently provides several other features, such as introduction of role abbreviations (to use in the spreadsheet and the graphic map), definition of sub-sets of functional roles (Figure 1, (1)), and the field for notes and comments. The concept of sub-sets plays an important role in subsystems encoding and interpretation. They usually represent the most compact units, such as multi-subunit complexes, or variants of pathways. Examples of sub-sets such as “Bacterial type” and “Eukaryotic type” are illustrated in (Figure 1, (2)). The current implementation of the SEED allows the reduction of a display to a selected sub-set of functional roles (Figure 1, (3)), and to a selected group of organisms (Figure 1, (4)).

2. Provide a graphic representation of a subsystem.

Graphic representation of a subsystem (a pathway map) is an important component of the expert analysis. The subsystem visualization in SEED currently relies largely on KEGG maps. These maps are hyperlinked both to the SEED annotations and to the whole rich content of KEGG, including annotations, enzyme descriptions, reactions and compounds. However, a critical constraint in using KEGG for our purposes is a lack of capability to edit, modify and define new subsystems. Subsystem components (enzymes, compounds, reactions), which are currently not represented in KEGG collection, can be illustrated by appended pages provided by the experts in a free format (exemplified in Figure 6). Based on our experience, this minimal approach to graphics provides an excellent versatile environment for subsystem analysis without deviating valuable resources on producing elaborate images, which would only slow down and decrease the performance of experts who usually operate in the environment with a limited bandwidth.

Column	Abbrev	Functional Role
1	TDO	Tryptophan 2,3-dioxygenase (EC 1.13.11.11)
2	IDO	Indoleamine 2,3-dioxygenase (EC 1.13.11.42)
3	KFA_e	Kynurenine formamidase (EC 3.5.1.9)
4	KFA_b	Kynurenine formamidase, bacterial (EC 3.5.1.9)
5	KMO	Kynurenine 3-monooxygenase (EC 1.14.13.9)
6	KYN	Kynureninase (EC 3.7.1.3)
7	HAD	3-hydroxyanthranilate 3,4-dioxygenase (EC 1.13.11.6)
8	ASPOX	L-aspartate oxidase (EC 1.4.3.16)
9	ASPDH	Aspartate dehydrogenase [same functional role as] (EC 1.4.3.16)
10	QSYN	Quinolate synthetase (EC 4.1.99.-)
11	QAPRT	Quinolate phosphoribosyltransferase [decarboxylating] (EC 2.4.2.19)
12	NAMNAT	Nicotinate-nucleotide adenylyltransferase (EC 2.7.7.18)
13	NMNAT	Nicotinamide-nucleotide adenylyltransferase (EC 2.7.7.1)
14	NADS	NAD synthetase (EC 6.3.1.5)
15	GAT	Glutamine amidotransferase chain of NAD synthetase
16	NADK	NAD kinase (EC 2.7.1.23)
17	NAM	Nicotinamidase (EC 3.5.1.19)
18	NAPRT	Nicotinate phosphoribosyltransferase (EC 2.4.2.11)
19	NMPRT	Nicotinamide phosphoribosyltransferase (EC 2.4.2.12)
20	PNUC	Ribosyl nicotinamide transporter, pnuC-like
21	RNK_b	Ribosyl nicotinamide kinase (EC 2.7.1.22)
22	RNK_e	Ribosyl nicotinamide kinase, eukaryotic (EC 2.7.1.22)

Figure 1: Subsystem “NAD and NADP cofactor biosynthesis”

A: Functional roles in subsystem

B: Subsystem construction/visualization controls:

Figure 2: Subsystem construction and analysis tools

1 ☐ fill

[Export subsystem data](#)

☐ ignore alternatives

☐ use external ids

☐ show clusters

☐ show missing

2 ☐ show missing with matches [To restrict to a single genome:

☐ check assignments [☐ strict] [To restrict to a single genome:

☐ refill spreadsheet from scratch

☐ show duplicates

☐ show PEGs in roles that do not match precisely

☐ Add Genomes with Solid Hits

☐ show coupled PEGs fast [depends on existing pins/clusters]

☐ show coupled PEGs [figure 2 minutes per PEG in spreadsheet]

Align column: ☐ show input to alignment tool

Include homologs that pass the following threshold:

Change annotation for column:

If you wish to check the subsystem, [click here](#)

You should resynch PEG connections only if you detect PEGs that show the code. If you find yourself having to use it, send mail to Ross.

3

3. Propagate a subsystem.

In the current implementation, genomes are added one-by-one as additional rows in the main spreadsheet. In the future SEED would support adding (and removing) genomes by groups selected on the phylogenetic tree or on the tree of homologs induced by a selected functional role (considered a *subsystem signature*). However, for photoautotrophs a beta-version of this tool has been already implemented that supports addition of their genomes as a group (see Figure 3A). Added genome rows are automatically populated by IDs of genes with matching annotations. These IDs contain embedded URL hyperlinks to individual gene pages (not shown) displaying details for each gene/protein, including genome environment, annotations, link to external databases, and provide access to similarity tables, genome context analysis and annotation tools. Once the subsystem has been manually populated with gene IDs for a set of model organisms it can be automatically propagated over all of the integrated genomes (Figure 2, (3)).

Figure 3. A: The SEED tool for the simultaneous subsystem propagation over any number of available photoautotrophic organisms. **B:** Tool to determine occurrence profiling

A

Pick Phototrophic Genomes to Extend with

Anabaena variabilis ATCC 29413 [B] (240292.1) [66 contigs]
Arabidopsis thaliana [E] (3702.1) [1675 contigs]
Crocospaera watsonii WH 8501 [B] (165597.1) [315 contigs]
Gloeobacter violaceus PCC 7421 [B] (251221.1) [1 contigs]
Nostoc punctiforme [B] (63737.1) [203 contigs]
Nostoc sp. PCC 7120 [B] (103690.1) [7 contigs]
Oryza sativa (japonica cultivar-group) [E] (39947.1) [303 contigs]
Prochlorococcus marinus MED4 [B] (167540.1) [1 contigs]
Prochlorococcus marinus str. MIT 9313 [B] (74547.1) [1 contigs]
Prochlorococcus marinus subsp. marinus str. CCMP1375 [B] (167539.1) [1 contigs]

☐ fill

update spreadsheet

B

Searching for Interesting Genes

Search for Genes Matching an Occurrence Profile or Common to a Set of Organisms

4. Propagate consistent annotations

SEED tool “Show missing with matches” (Figure 2, (3)) allows to request a list of functional roles in each organism for which no genes are yet connected to a spreadsheet. If requested, also best candidates for each role are automatically suggested from the tables of pre-computed homologs. The spreadsheet form makes propagation of annotations significantly less error prone by making it vividly clear which genes are very likely to exist in an organism, even if they have been miss-annotated or unidentified yet. It also aids in correction of “over-annotation” errors and in refinement of paralogs. Corrections of this type are supported by automated flagging of duplicate entries. Such entries may reflect genuine bi-functional proteins, but often reflect a propagation of annotation errors due to occasional protein fusion events, a notorious problem of annotation pipelines. The special color-coding is developed to flag outliers in sets of homologs.

5. Establish functional variants

Absence of a functional role (or a sub-set of roles) may reflect a realization of a different, albeit totally functional, variant of a system. We distinguish two types of variants associated with:

Figure 4: NAD and NADP cofactor biosynthesis: subsystem spreadsheet

Presence of genes assigned with specific functions is shown by SEED gene/protein IDs. Highlighting by a matching color indicates proximity on the chromosome. Subsets of functional roles corresponding to subsystem diagram (Figure 5) are shown.

Organism	Variant Code	ASPOX	ASPDH	QSYN	QAPRT	NAMNA T	NMNAT	NADS	GAT	NADK	NAM	NAPR T	NMPRT	PNUC	RNK_b
Prochlorococcus marinus str. MIT 9313 [B]	1	177		661	2064	1454		1455	1455	2008 368	niacine salvage			N- ribosyl- nicotin- amide salvage	
Synechococcus elongatus PCC 7942 [B]	1	de novo pathway				universal pathway									
Synechococcus sp. WH 8102 [B]	1	2185		1311	2311	500		499	499	1594 2261					
Prochlorococcus marinus MED4 [B]	4	350		1670	259	727		726	726	292 957					
Prochlorococcus marinus subsp. marinus str. CCMP1375 [B]	1	119		1011	213	1595		1596	1596	1340 180					
Prochlorococcus marinus subsp. pastoris str. CCMP1986 [B]	4	100		675	188	1441		1442	1442	1265 156					
Nostoc sp. PCC 7120 [B]	1	1527		4980	2398	2790 5370		2792	2792	5058 538	5469	2789			
Thermosynechococcus elongatus BP-1 [B]	1	2407		231	1712	1177		1179	1179	483 857	541 , 542 , 545	1176			
Gloeobacter violaceus PCC 7421 [B]	0	846		192	888	4016	1660	2587	2587	3525 473					
Synechocystis sp. PCC 6803 [B]	1	2958		2974	1816	694	2723	1807	1807	1443 1953			2724		
Organism	Variant Code	ASPOX	ASPDH	QSYN	QAPRT	NAMNA T	NMNAT	NADS	GAT	NADK	NAM	NAPR T	NMPRT	PNUC	RNK_b
Nostoc punctiforme [B]	1	4768		6260	6375	5019 5605	3160	5607	5607	981	1255	5604		3159	3160

Figure 5:
NAD and NADP biosynthesis in cyanobacteria: subsystem diagram

A. known non-orthologous gene displacements, and
A. known or inferred alternative metabolic routes (fluxes)

A. The first term reflects a situation when the same (or metabolically equivalent) functional role in various species is encoded by non-homologous or highly divergent genes. With the growing number of sequenced genomes we witness more and more examples of this kind (for a review see {Galperin, 1999 #4}). For example, in the NAD(P) biosynthesis the first step is driven by L-aspartate oxidase (ASPOX) in most bacteria (Figure 4). However, *T. maritima* and several anaerobic archaea have an alternative enzyme: aspartate dehydrogenase (ASPDH).

B: Encoding of functional variants representing alternative fluxes. Decision of what constitutes a functional variant depends on the type of a subsystem and its overall biological goal. In the illustrated case, the production of NAD is a well-defined goal, which allows to access functional variants with respect to this goal. Briefly, the following functional variants of NAD biosynthesis can be distinguished in cyanobacteria (see Figures 4 and 5). A union of the *de novo* and universal pathways builds a complete path from L-aspartate to NAD(P). This variant (#1) reflects purely autotrophic lifestyle characteristic of such organisms as *Synechococcus* sp. WH8102 and *Prochlorococci*. This is in concord with their habitat - oligotrophic oceans. On the contrary, genome of *Nostoc punctiforme* is the richest in NAD salvage pathways (variant 2). This is consistent with a perceived availability of various NAD precursors in the unique ecological niche of *N. punctiforme* - fungal endocyanosis *Geosiphon pyriforme*. A single (niacine) salvage pathway is present in genomes of *Nostoc* sp. PCC 7120 and *T. elongates* – meso/eutrophic freshwater species (variant 3). *Synechocystis* 6803 harbors a rare two-step “NMPRT-NMNAT” NAD biosynthesis from nicotinamide via NMN in addition to the universal NaMNAT-dependent pathway (variant 4). This and multiple other predictions regarding various aspects of NAD and NADP biosynthesis in cyanobacteria have been experimentally verified based on *in vitro* characterization of several key enzymes of NAD biosynthesis (NaMN adenyltransferase, nicotinamide phosphoribosyltransferase, and NAD synthetase), and on genetic analysis of the NMNAT and NaMNAT knockout mutants in *Synechocystis* 6803. Results of this comprehensive analysis are been prepared for publication – see manuscript summary below.

6. Locate missing genes

One of the most interesting outcomes of Subsystem-based metabolic reconstruction is the prediction of “missing genes” encoding functional roles inferred by the analysis of subsystems but not connected to specific genes in all or a sub-set of species. The concept of the “missing gene” has been recently reviewed in (Osterman and Overbeek, 2003). Based on our experience and preliminary analysis of selected sub-systems, we anticipate that a significant number of “missing gene” cases will be revealed by a detailed metabolic reconstruction effort in any microorganism. By our estimates, more than 10% of all enzymatic functions reliably characterized by classic techniques have never been connected to specific genes in any organism (“globally missing genes”). The actual fraction of missing genes in a given sub-set of organisms may be much higher, due to the nonorthologous gene displacement, as briefly discussed above (“locally missing genes”).

For example, 4-vinyl-protochlorophyllide reductase, (EC 1.3.1.75), an enzyme involved in chlorophyll biosynthesis remains a missing gene in cyanobacteria and higher plants. The gene *bchJ* encoding this enzyme in *Rhodobacter* sp. has no orthologs in other groups of chlorophyll-containing organisms. Among locally missing genes involved in pigment biosynthesis in cyanobacteria are Chlorophyll/chlorophyllide an oxygenase (EC 1.14.13.-) and Lycopene beta cyclase (EC 1.14.-.-) in *Prochlorococci*, several lysine biosynthesis enzymes in all cyanobacteria, etc. **Table 1** in the main text of this Report lists many of the “missing genes” revealed in

cyanobacterial genomes by SEED subsystem curators, some with inferred gene candidates. The complete and ever-growing log of missing genes and predictions and conjectures (where available) is maintained in the Subsystems Forum (<http://brucella.uchicago.edu/SubsystemForum/>), which can be accessed from CyanoSEED WEB portal (<http://cyanoseed.thefig.info/>).

7. Generate functional conjectures

Filling-in missing pieces in metabolism is a fundamentally important task. In the context of the integrative study of energy, carbon, and redox metabolism in *Synechocystis* 6803 this task is of special importance. The identification of novel genes encoding enzymatic functions inferred by subsystem analysis, is also required to validate functional variants of subsystems. The type of analysis performed at this stage is illustrated in Appendix II (using Carotenoids biosynthesis subsystem as an example) and in the summary of our manuscript on NAD and NADP biosynthesis in cyanobacteria.

Summary of the manuscript in preparation:

“Genomics of NAD biosynthesis and recycling in cyanobacteria”

1. NAD and NADP are essential cofactors in all forms of life. In accordance with this, genes coding for enzymes of the universal NAD/NADP biosynthetic pathway (Figure 4) can be asserted in all cyanobacteria. In some species these genes tend to occur in chromosomal clusters. This aspect of genome context (proximity on a chromosome) is highlighted by matching colors in Figure 4. Such clustering allowed prediction and verification of *sll1916* as the gene encoding NaMNAT in *Synechocystis* 6803 (Kurnasov et al., 2005).
2. NaMN, the major precursor of NAD can be synthesized *de novo* or obtained by salvage of several exogenous compounds. Subsystem analysis allowed us to infer that *Synechococcus* sp. WH8102 and *Prochlorococci* carry out a complete *de novo* pathway, but are incapable of NAD salvage, which is in concord with their habitat - oligotrophic oceans. On the contrary, genome of *Nostoc punctiforme* is the richest in NAD salvage pathways (Figure 4). This is consistent with a perceived availability of various NAD precursors in the unique ecological niche of *N. punctiforme* - fungal endocyanosis *Geosiphon pyriforme*.
3. Unlike all other cyanobacteria, *Synechocystis* 6803 harbors a rare two-step “NMPRT-NMNAT” NAD biosynthesis from nicotinamide via NMN (Figures 4 and 5) in addition to the universal NaMNAT-dependent pathway. We have predicted that NMNAT-dependent route is involved in NAD recycling, while the NaMNAT-depending route is responsible for *de novo* NAD production. This conjecture was confirmed by genetic techniques, monitoring of nicotinamide uptake and enzymatic analysis of previously uncharacterized NaMNAT and NMPRT of *Synechocystis* 6803 (Kurnasov et al., 2005).
4. A possible involvement of this pathway in salvage of the exogenous niacinamide by *Synechocystis* 6803 has been excluded experimentally. Therefore, NMPRT-NMNAT shunt in this organism is likely to be associated with recycling of NAD precursors formed during intracellular NAD degradation due in part to enhanced DNA repair (Emanuelli, 2001). The presence of an additional NAD recycling pathway in cyanobacteria can be indeed beneficial: elevated levels of DNA damage are likely to occur in these organisms as a byproduct of oxygenic photosynthesis.
5. Similarity searches in SEED detected simultaneous presence of NMPRT and NMNAT orthologs in genomes of three wide-ranging T4-type bacteriophages KVP40, 44RR2.8t, and Aeh1. The co-occurrence in a phage genomes of these genes (Figure 6) possibly conveying

Figure 6. Conserved chromosomal clustering of the genes encoding (1) Nicotinamide phosphoribosyltransferase (NMPRT) and (2) Nicotinamide-nucleotide adenyltransferase (NMNAT) / ADP-ribose pyrophosphatase in the genomes of various bacteria and a bacteriophage 44RR2.8t. Alignment of chromosomal contigs of respective genomes around NMPRT. Homologous genes are shown by arrows with matching colors and numbers. Genes not conserved within the clustered are colored pink.

selective advantages to its host is a rare example of phage potential for mediating horizontal transfer of a complete metabolic pathway.

6. Notably a combination of genome and functional context analysis supported by SEED helps to improve gene annotations – even in draft genomes: eg. *S. elongates* is likely to possess the NMNAT and NMPRT orthologs (Figure 6).

References

1. Osterman, A. and Overbeek, R., Missing genes in metabolic pathways: a comparative genomics approach. *Curr Opin Chem Biol* 2003, 7, 238-51.
2. Galperin, M. Y. and Koonin, E. V., Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes. *Genetica* 1999, 106, 159-70.
3. O. Kurnasov, S. Gerdes, K. Shatalin, B. Polanuyer, and A. L. Osterman. (2005) Genomics of NAD biosynthesis and recycling in cyanobacteria. *Mol. Microbiol.*, manuscript in preparation.

Appendix II.

Example #1: Electron transfer in *Synechocystis* 6803.

Photosynthesis, respiration, and other central metabolic functions are often treated separately, but in a prokaryotic system these occur in the same cell and often in the same membrane. Indeed, in cyanobacteria the photosynthetic and respiratory electron transfer pathways co-occur in the thylakoid membrane with both electron transfer pathways intersecting at the plastoquinone (PQ) pool and the cytochrome b6/f complex. Thus, PQ redox state could be a very informative probe of cell physiology under different environmental and genetic conditions. The lab of Dr. Vermaas, our collaborators on this project, have designed a method by which the PQ redox state in the thylakoids can be determined in vivo (Howitt et al., 2001) using a quinone (Q) electrode (Dry et al., 1989; Zannoni and Moore, 1990; Cleland, 1998). To facilitate interpretation of complex and often unexpected results of these PQ redox state measurements, the FIG team is working on identification and cataloging (based on genomic data) all cellular sources and “sinks” of electrons potentially functional in *Synechocystis* 6803. This is an interactive process spiraling along the path of genome analysis -> functional assertions -> experimental data -> model refinement -> experimental verification... The current state of this work is summarized in Figure 1 and is encoded in the following subsystems available on CyanoSEED at (<http://cyanoseed.thefig.info/>):

Level I	Level I	Level I	Subsystems
Oxidative Phosphorylation	Electron donating reactions		NADH-quinone oxidoreductase
			NADH-ubiquinone oxidoreductase
			Succinate dehydrogenase
			Hydrogenases
			Formate hydrogenase
			Anaerobic_respiratory_dehydrogenases
			Aerobic_respiratory_dehydrogenases
		Quinone pool reductase	Cytochrome B6-F complex
			Ubiquinone Menaquinone-cytochrome c reductase complexes
		Electron accepting reactions	Terminal cytochrome C oxidases
			Terminal cytochrome oxidases
			Anaerobic respiratory reductases
			Nitrate,nitrite,nitrous oxide reductases
			Fumarate reductase
		Soluble electron carries	
			Soluble_cytochromes and functionally related electron carrier
		ATP synthases	
			F0F1-type ATP synthase
			V-Type ATP synthase
		Soluble electron carries	
		Transhydrogenation	
		Protection from reactive oxygen species	
Photosynthesis	Oxygenic photosynthesis	Electron transfer	Photosystem I
			Photosystem II
			Cytochrome B6-F complex
		Light harvesting	Phycobilisome

Each Subsystem name is hyperlinked to its actual encoding in SEED.

Figure 1. Electron transfer in *Synechocystis* 6803 and other cyanobacteria – an overview. Enzymes and protein complexes donating electrons to PQ pools are on the left, potential “electron sinks” are on the right. The presence of genes with respective functions in the genome of *Synechocystis* 6803 is shown by green shading.

The group of Dr. Vermaas has found succinate dehydrogenase (SDH) to be the major respiratory electron donor to the PQ pool (Cooley et al., 2000). This unexpected prominence of SDH in respiratory electron transfer has been backed up by additional experimental evidence (Cooley and Vermaas, 2001). Furthermore, Type 1 NADPH dehydrogenase (NDH-1; the Complex I equivalent in cyanobacteria), until recently considered to be the sole respiratory electron carrier into the thylakoid PQ pool in cyanobacteria, has a low capacity of electron donation to the PQ pool in the thylakoid. To support this study, the FIG team has analyzed *Synechocystis* 6803 succinate dehydrogenase (SDH) *in silico*. The SDH generally works as a complex of flavoprotein, iron-sulphur and cytochrome subunits. In many organisms the fourth, membrane anchor subunit is also present. Surprisingly, all cyanobacteria, including *Synechocystis* 6803, were found to lack the common form of Succinate dehydrogenase cytochrome b-556 subunit (Figure 2 (1)). A candidate gene for this functional role has been identified by long-range similarity searches in the genomes of *Prochlorococcus* and *Synechococcus* WH 8102 (pmt0131 for example, Figure 2 (2)), but not in any other cyanobacteria. As for *Synechocystis* 6803 (as well as *T. elongatus*, *A. variabilis*, *Nostoc* PCC 7120, *N. punctiforme*, *C. watsonii*, and *T. erythraeum*) -- we predict that these organisms harbor orthologs of an unusual archaeal SdhC subunit (slr0201 in SY, Figure 2 (3)), which lacks typical transmembrane alpha-helical regions and exhibits remarkable 30% sequence similarity to the heterodisulfide reductase (Janssen et al, 1997). Cyanobacteria lack Succinate dehydrogenase anchor subunit SdhD as well. Search for the candidate genes is in progress.

Organism	Variant Code	SDC	SDIS	SDFAD	SDM	SDC2	pSDC	HSDM	SDMN
<i>Prochlorococcus marinus</i> MED4	0								
<i>Prochlorococcus marinus</i> subsp. <i>pastoris</i> str. CCMP1986	0								
<i>Synechocystis</i> sp. PCC 6803	0		slr1625 slr0823	slr1233			slr0201		
<i>Anabaena variabilis</i> ATCC 29413	0		4686	4956			5984		
<i>Nostoc</i> sp. PCC 7120	0		1255	3277			3648		
<i>Crocospaera watsonii</i> WH 8501	0		4905	456			3763		
<i>Gloeobacter violaceus</i> PCC 7421	0		2944	2988			2945		
<i>Thermosynechococcus elongatus</i> BP-1	0		1753	1376			978		
<i>Trichodesmium erythraeum</i> IMS101	0		6019	1251			4571		
<i>Nostoc punctiforme</i>	0		1073	2879			6352		702
Organism	Variant Code	SDC	SDIS	SDFAD	SDM	SDC2	pSDC	HSDM	SDMN
<i>Prochlorococcus marinus</i> str. MIT 9313	0		133	132		131			
<i>Prochlorococcus marinus</i> subsp. <i>marinus</i> str. CCMP1375	0		698	697		696			
<i>Synechococcus elongatus</i> PCC 7942	0		857	1931		2229			
<i>Synechococcus</i> sp. WH 8102	0		585	586		587			

Figure 2. Succinate Dehydrogenase: subsystem spreadsheet. The presence of genes assigned with functional roles (see the list of abbreviations below) in the *Synechocystis* 6803 genome is shown by CyanoBase IDs, in all other genomes – by SEED IDs. Highlighting indicates proximity of genes on a chromosome.

Abbrev	Functional Role
SDC	Succinate dehydrogenase cytochrome b-556 subunit (EC 1.3.99.1)
SDIS	Succinate dehydrogenase iron-sulfur protein (EC 1.3.99.1)
SDFAD	Succinate dehydrogenase flavoprotein subunit (EC 1.3.99.1)
SDM	Succinate dehydrogenase hydrophobic membrane anchor protein
SDC2	Succinate dehydrogenase cytochrome b subunit
pSDC	putative succinate dehydrogenase subunit C (EC 1.3.99.1)
HSDM	Hypothetical succinate dehydrogenase membrane anchor protein
SDMN	Succinate dehydrogenase hydrophobic membrane anchor protein(Nostoc p.)

Example #2: Subsystem Carotenoids Biosynthesis

Carotenoids are synthesized by plants, photosynthetic bacteria (carotenes, xanthins), and archaea (bacteriorhodopsin) as components of light harvesting complexes and UV-protectors, as well as by non-photosynthetic bacteria, where they function mainly as UV-protectors. Carotenoids differ by the length of carbon skeleton (C-30, C-40, C-50), the number and type of cyclic structures in a molecule (one or two rings of beta or epsilon conformation), the presence/absence of ketogroups, and by nature of carbohydrate side chains. About 30 enzymes involved in biosynthesis of various carotenoids are presently known. However, the biochemistry of many stages of carotenoid biosynthesis is still a mystery. In addition, genes encoding several known enzymes are unknown in many organisms, including the cyanobacteria ("missing genes"). See Figures 1 and 2 for several examples: (i) orthologs of lycopene beta cyclase (**BLCY**), although present in genomes of *Prochlorococci*, can not be identified via homology projections in (complete) genomes of *Synechocystis* 6803, *T. elongates*, *Nostoc* sp PCC 7120, and several other cyanobacteria; (ii) orthologs of beta-carotene ketolase (**BCK**), the first enzyme of the branch leading to ketocarotenoids (Figure 2), are missing in the genomes of *T. elongates*, *C. watsonii*, *T. erythraeum*, *S. elongates* and several *Prochlorococci*; (iii) orthologs of several enzymes leading to myxoxanthophyll, including hydroxyneurosporene synthase (**HNS**) and hydroxyneurosporene methyltransferase (**HNM**) are unknown in any organism ("globally missing").

Table 1: Functional roles in Subsystem. A subset of roles is listed: known enzymes of carotenoid biosynthesis in cyanobacteria (abbreviations correspond to those in Fig. 1, 2)

PS	Phytoene synthase (EC 2.5.1.32)
PDS1	Phytoene desaturase, pro-zeta-carotene producing (EC 1.-.-.-)
NDS	Neurosporene desaturase (EC 1.-.-.-)
GPS	Geranylgeranyl pyrophosphate synthase (EC 2.5.1.29)
CI	Carotenoid cis-trans isomerase (EC 5.2.-.-) ,
ZDS	Pro-zeta-carotene desaturase, polycopene producing (EC 1.-.-.-)
BCH	Beta-carotene hydroxylase (EC 1.14.13.-)
BLCY	Lycopene beta cyclase (EC 1.14.-.-)
BCK	Beta-carotene ketolase (EC 1.14.-.-)
ELCY	Lycopene epsilon cyclase
ZGT	Zeaxanthin glucosyl transferase
HNS	Hydroxyneurosporene synthase
HNM	Hydroxyneurosporene methyltransferase (EC 2.1.1.-)
PDSB	Phytoene desaturase, neurosporene or lycopene producing (EC 1.3.-.-)

Functional roles branching to pigment biosynthesis in non-photosynthetic bacteria

DSS	Dehydrosqualene synthase (EC 2.5.1.-) ,
SM	Spheroidene/spirilloxanthin monooxygenase
LE	Lycopene elongase (EC 2.5.1.-)
PCS	"phi-Carotenoid synthase" (EC 1.3.-.- and EC 2.1.1.-)

Figure 1: Subsystem Spreadsheet. The presence of genes assigned with functional roles (see the list of abbreviations above) is shown by SEED IDs.

Organism	Variant Code	PS	PDS	NDS	GPS	CI	ZDS	BCH	BLCY	BCK	ELCY	ZGT	HNS	HNM
Gloeobacter violaceus PCC 7421	0	1744			416. 753	2133			2874	1728. 394				
Prochlorococcus marinus subsp. pastoris str. CCMP1986	0	143	1562		1066	1111	115		1060		630			
Synechococcus sp. WH 8102	0	2247	2248		737	896	2206	288	723					
Prochlorococcus marinus MED4	0	305	607		1185	1138	333	212	1191					
Prochlorococcus marinus subsp. marinus str. CCMP1375	0	166	167		1127	583	136	266	1134		788			
Prochlorococcus marinus str. MIT 9313	0	1995	1996		1107	1049	1960. 953	1810	1121	1552	1767			
Thermosynechococcus elongatus BP-1	0	1559	1560	232	20		336	1899						
Crocospaera watsonii WH 8501	0	762	761	3206	2154	2857	2233. 4102	1038. 4133				411		
Trichodesmium erythraeum IMS101	0	5153	5154	3089	1721	6180	5085	5965				6281		
Synechococcus elongatus PCC 7942	0	1293	1292	313	125	577	836	1734			1370			
Organism	Variant Code	PS	PDS	NDS	GPS	CI	ZDS	BCH	BLCY	BCK	ELCY	ZGT	HNS	HNM
Nostoc sp. PCC 7120	0	2140	2139	5430	524	2371	2095. 2689	4316		4051				
Anabaena variabilis ATCC 29413	0	3722	3721	4341	5421	4633	4853. 6105	1577		1964				
Synechocystis sp. PCC 6803	0	1269	1268	261	111. 3167	2828	1821	889		2592		73		
Nostoc punctiforme	0	2875	2874	5675	5170. 745	3633	464. 859	5347		184. 772		2699		
Chlorobium tepidum TLS	0	1367	795		254	640	1394		180				299	28

Figure 2: Subsystem diagram
(see legend below)

Figure legend: Subsystem diagram showing enzymes (boxes with abbreviations matching Table 1), key reaction intermediates (circles with roman numerals), and reactions (arrows). The presence or absence of genes with respective functions in the genomes of various cyanobacteria are shown using color-coded shading as explained in the panel. Abbreviations of reaction intermediates are as follows: I-Farnesyl pyrophosphate, I-Geranylgeranyl pyrophosphate, II- Phytoene, III-all-trans zeta-Carotene, IV-trans-Lycopene, V- neurosporene, VI-3,4-didehydroneurosporene, VII-3,4-didehydrolycopene, VIII-torulene, IX-deoxymyxol, X-Deoxymyxol 2'-(2,4-di-O-methyl-alpha-L-fucoside, XI-Deoxymyxol 2'-(2,4-di-O-methyl-alpha-L-fucoside, XII-ihydroxyneurosporene, XIII-trans Lycopene, XIV-Zeaxanthine, XV-beta-Carotene, XVIcis zeta-carotene, XVIIcis Lycopene, XIXgamma-Carotene, XXalpha-Carotene, XXIdelta-Carotene, XXIIlutein, XXIII-astaxanthin, XXIV-canthaxanthin, XXV-beta-isorenieratene, XXVI-Echinenone, XXVII-3'-Hydroxyechinenone, XXVIII-3,4-dihydrodeoxymyxol, XXIX-3,4-dihydromyxol, XXX-3,4-dihydromyxol mono o di-methylfucoside, XXXI-diapophytoene, XXXII-retinal

The FIG team was able to suggest candidate genes for lycopene beta cyclase (and for other inferred functional roles – see Table 2 in the Report) in *Synechocystis* 6803 and other experimentally tractable cyanobacteria based on the combination of advanced techniques of the genome context and functional context analysis supported by SEED. This type of analysis is illustrated below.

First, gene candidates for the role of lycopene beta cyclase (**BLCY**) were selected on the basis of gene clustering to known members of carotene biosynthesis pathway (7 candidates) or sequence similarity to known forms of lycopene cyclase (6 candidates). Hypothetical proteins as well as hypothetical cyclases, synthases, and dehydrogenases were all considered in this preliminary screen. However, the search favored proteins with evident transmembrane helices, since known lycopene cyclases are membrane proteins. Occurrence Profile (the presence of a strong ortholog in each cyanobacterial genomes where lycopene cyclase is “missing”, and its absence from each genome where lycopene cyclases is evident) was used as an additional filter. Next, each of the preliminary candidate ORFs was analyzed in detail. The top five candidates selected in *Synechocystis* 6803 are listed in Table 1 and discussed below:

Candidate #1: Hypothetical protein Slr0941

Supporting evidences: (i) homologous to oligoketide cyclase/lipid transport protein, (ii) is a membrane protein, (iii) very strong chromosomal clustering: is co-localized with orthologs of pro-zeta-carotene desaturase, prolycopene producing (**ZDS**) in *Synechocystis* 6803, as well as in nearly all other analyzed cyanobacterial genomes (see Figure 1 below).

Negative features: (i) orthologs are present in *Prochlorococci*, but (ii) absent in the genome of *Chlorobium tepidum*.

Candidate #2: Two closely related hypothetical dehydrogenases Sll0659 and Sll0147:

Supporting evidences: (i) nearly perfect occurrence Occurrence Profile, (ii) proposed function – dehydrogenase, (iii) presence a UbiH domain (Ubiquinone biosynthesis reductase), hence potential activity towards related isoprenoid derivatives, (iv) distant homology with lycopene cyclase (CrtL) from *Prochlorococci*.

Table 1: Candidate ORFs for the missing lycopene cyclase. Orthologs in *Synechocystis* 6803 genome are identified by their CyanoBase IDs, in all other genomes – by GenBank IDs. Grey boxes indicate the absence of an ortholog in a given genome or the presence of a very weak one. Close orthologs identified by reliable homology projections are in yellow boxes.

Organism/protein	I	II	III	IV	V
<i>Synechocystis</i> sp. PCC 6803	slr0941	sll0659/ sll0147	sll0513	slr1579	slr1394
<i>Chlorobium tepidum</i> TLS		gi 21673292			
<i>Thermosynechococcus elongatus</i> BP-1	gi 22297879	gi 22298682	gi 22298640	gi 22299456	gi 22297563
<i>Synechococcus</i> sp. WH 8102	gi 33866744			gi 33866807	gi 33865277
<i>Prochlorococcus marinus</i> MED4	<50% homology				
<i>Prochlorococcus marinus</i> str. MIT 9313	gi 33864232				gi 33863379 <50% homol
<i>Prochlorococcus marinus</i> CCMP1375	gi 33239587			gi 33239638 <18% homol	gi 33240578
<i>Synechococcus elongatus</i> PCC 7942	gi 46129933	gi 45511722		gi 45511757	gi 46129485
<i>Nostoc</i> sp. PCC 7120	gi 17229873	gi 17228415 gi 17231016	gi 17229297	gi 17231534	gi 172277
<i>Nostoc punctiforme</i>	gi 23130102	gi 23126269 gi 23130367	gi 23125690	gi 23127750	gi 23129765
<i>Anabaena variabilis</i> ATCC 29413	peg.6104	gi 46135210	Alr1805	gi 45508370	gi 45510320
<i>Crocospaera</i> w, WH8501	ref ZP_00178762.1	gi 46118345 gi 45528290	gi 46119198	gi 46119504	gi 46118491
<i>Trihodesmium</i> IMS101	gi 48893657	gi 48894952	gi 48891505	gi 48890958	gi 48893862

Candidate #3: Hypothetical protein Sll0513

Supporting evidences: (i) highly homologous to Phytoene/squalene synthetase and/or Farnesyl-diphosphate farnesyl-transferase (EC2.5.1.21), (ii) distant homology to lycopene cyclase / phytoene synthase from *Blakeslea trispora*, (iii) satisfactory Occurrence Profile, (iv) is a membrane protein.

Negative feature: absent in the genomes of *Chlorobium tepidum* and *T. elongates*

Candidate #4: Hypothetical protein Slr1579

Supporting evidence: (i) homology, albeit distant, with CrtYm (monocyclic)-lycopene cyclase, functionally analogous to the one inferred in *Synechocystis* 6803 (monocyclic lycopene cyclase). This low homology may be still meaningful, since CrtYm has the same low homology to *Prochlorococcus* lycopene cyclase (CrtY)

Negative feature: absent in *Chlorobium tepidum*

Candidate #5: Hypothetical protein Slr1394

Supporting evidence: (i) strong chromosomal clustering with Geranylgeranyl pyrophosphate synthase (GPS) in cyanobacteria (excluding *Synechocystis* 6803), (ii) integral membrane protein, (iii) orthologs are evident in plants, (iv) satisfactory occurrence.

Negative features: ortholog is absent in *Chlorobium tepidum*, and present (rather distant however) in *Prochlorococci* and *Synechococcus* sp. WH 8102.

Figure 1. Clustering on the chromosome of one of the gene candidates for the role of lycopene beta cyclase (**BLCY**) with orthologs of pro-zeta-carotene desaturase, prolycopene producing (**ZDS**) in nearly all cyanobacterial genomes available to date. Alignment of chromosomal contigs of respective genomes around the query ORF, slr0941. Homologous genes are shown by arrows with matching colors and numbers as follows: (1) lycopene beta cyclase (**BLCY**), (2) pro-zeta-carotene desaturase, prolycopene producing (**ZDS**). Genes not conserved within the clustered are colored pink.

Example #3: Methionine biosynthesis in cyanobacteria: open questions

Methionine and cysteine are the two sulfur-containing amino acids. In addition to its general function as a component of proteins, methionine is specifically required for translation initiation and is crucial for a variety of methyltransferase reactions as a precursor of S-Adenosyl-Methionine (SAM). Main stages of methionine biosynthetic pathway (Figure. 1) are:

Homoserine is derived from aspartate semialdehyde by the homoserine dehydrogenase (HSDH). Acylation of homoserine is catalyzed either by homoserine succinyltransferase (HSST) or by homoserine acetyltransferase (HSAT), unrelated to HSST. Yet another variation in homoserine esterification exists in plants, where O-phosphoryl-L-homoserine is produced in a reaction catalyzed by homoserine kinase (HK). Cysteine serves as a precursor for methionine biosynthesis via the trans-sulfuration pathway mediated by two enzymes, cystathionine gamma-synthase (CTGS) and cystathionine beta-lyase (CTBL). An alternative pathway for methionine biosynthesis, the sulfhydrylation pathway, utilizes inorganic sulfur instead of cysteine and is catalyzed by O-acetylhomoserine sulfhydrylase (AHS). Both pathways can utilize O-succinyl-L-homoserine or O-acetyl-L-homoserine in different bacterial species. The methylation of homocysteine by methyl-THF in bacteria can be catalyzed by two types of methionine synthases. Reaction catalyzed by coenzyme B12-dependent protein MetH is more than 100-fold faster than the reaction catalyzed by B12-independent isoenzyme MetE. In many bacteria the methyl group of methionine is donated by methyl-THF, which is formed by reduction of methylene-THF in a reaction catalyzed by MTHFR. S-adenosylmethionine (SAM) is synthesized from methionine and ATP by SAM synthetase. Utilization of SAM as a methyl donor results in formation of S-adenosylhomocysteine (SAH), which is then cleaved to homocysteine and adenosine by SAH hydrolase (Fig. 1).

Remarkably, multiple variations of methionine biosynthesis in its flow from homoserine to homocysteine exist in different cyanobacteria. *Prochlorococcus* and *Synechococcus* WH 8102, containing orthologs of HSST and AHS (clustered on a chromosome), as well as CTGS and CTBL (clustered as well) apparently possess both pathways, each utilizing O-succinyl-L-homoserine (HSST is asserted, while HSAT is not) (**Variant 1**). Based on the presence of HSAT and AHS homologs in their genomes, *A. variabilis* and *N. punctiforme* utilize only sulfhydrylation pathway with succinyl-CoA as precursor for homoserine esterification (**Variant 2**). The sulfhydrylation pathway in *S. elongatus* apparently utilizes O-acetyl-L-homoserine (the clustered HSST and AHS orthologs can be asserted in its genome, **Variant 3**). Surprisingly, none of these pathway variations can be asserted in *Synechocystis* 6803, *Thermosynechococcus elongatus*, and several other cyanobacterial species (**Variant 4**). Since orthologs of homoserine dehydrogenase, as well as all enzymes catalyzing conversions from homocysteine to methionine are present, these organisms are expected to possess a functional route of L-homocysteine biosynthesis from homoserine. One possibility is that these cyanobacteria harbor a completely novel pathway of homocysteine biosynthesis, yet to be discovered. Alternatively, they may follow the plant route through O-phosphoryl-L-homoserine (orthologs of homoserine kinase are evident in all cyanobacterial genomes). However, homologs of cystathionine gamma-synthase (CTGS) and cystathionine beta-lyase (CTBL), catalyzing these conversions in plants are “missing” from these cyanobacterial genomes and are likely to be encoded by non-orthologous genes. The search for the gene candidates for the “missing” enzymes is in progress and is based on the following assumptions:

1. strong gene candidates are expected to be present mainly in the cyanobacterial genomes missing the known pathways of homocysteine biosynthesis, and absent in the organisms where other pathways are evident (**occurrence profile**);
2. methionine-specific regulatory sites are likely to be present upstream of the candidate genes (**coregulation**);
3. candidate genes are likely to be adjacent on a chromosome - based on a strong pattern of orthologs clustering in the known pathways of homocysteine biosynthesis (**clustering on a chromosome**).

Methionine Biosynthesis and SAM Recycling in cyanobacteria.

A: Subsystem spreadsheet

B: Subsystem diagram:

Organism	HSDH	HSST	HSAT	AHSH	CTGS	CTBL	MetH	MetE	MTHFR	SAMS	AHMC
<i>Crocospaera watsonii</i> WH 8501	3291							5009	4239	4094	1914
<i>Gloeobacter violaceus</i> PCC 7421	4295						477		789	2577	3183
<i>Synechocystis</i> sp. PCC 6803	2356						2469				
<i>Thermosynechococcus elongatus</i> BP-1	277						2473	1090	1144	308	1500
<i>Synechococcus elongatus</i> PCC 7942	1397	2172		2173			1027		1770	977	2389
<i>Anabaena variabilis</i> ATCC 29413	2331		3872	3873			702		639	1756	1953
<i>Nostoc punctiforme</i>	2895		5301	5302			4254		6434	2194	
<i>Nostoc</i> sp. PCC 7120	4427						6365		1885	2334	
<i>Trichodesmium erythraeum</i> IMS101	4266						4055		1093	3551	1724
<i>Synechococcus</i> sp. WH 8102	706	845		846	669	669	619		106	4431	
<i>Prochlorococcus marinus</i> str. MIT 9313	1141	875		874	225	225	1229	2279	4433	4394	1091
<i>Prochlorococcus marinus</i> subsp. <i>marinus</i> str. CCMP1375	1148	799		798	404	404	1233		2258	1980	116
<i>Prochlorococcus marinus</i> subsp. <i>pastoris</i> str. CCMP1986	1047	640		639	405	405	728		2005	1664	138
<i>Prochlorococcus marinus</i> MED4 [B]	1204	1714		1715	1	1	957		176	349	1783
							874		153	309	1621
							1421		295	134	548

Example 4: Quinone biosynthesis in cyanobacteria: open questions

Synechocystis 6803 and other cyanobacteria are known to contain plastoquinone and phyloquinone, and lack ubiquinone. The only known function for menaquinone in cyanobacteria is to provide a precursor for phyloquinone biosynthesis. In good agreement with this, a complete set of functional roles for menaquinone biosynthesis, as well as S-adenosylmethionine:2-demethylmenaquinone methyltransferase (MenG) converting menaquinone to phyloquinone can be asserted in all cyanobacterial genomes analyzed (Figure 1). A few omissions observed in draft genomes are likely due to sequencing gaps. Interestingly, a nearly complete set of orthologs of the ubiquinone biosynthesis genes is apparent in all cyanobacterial genomes (Figure 2), in spite of the absence of ubiquinone in these organisms. The overwhelming co-occurrence of all these genes and the fact that this pathway is largely preserved even in the “minimal” *Prochlorococcus* genomes argues for its physiological significance in cyanobacteria. Its function is a mystery. The possibility for these genes to be involved in plastoquinone (and not ubiquinone) biosynthesis has been proposed in (Dufresne, et al. 2003). However, plant-type plastoquinone biosynthetic pathway from homogentisate can be asserted in all cyanobacteria (with two “missing genes” in several cyanobacterial genomes - Figure 5).

The notable absence of chorismate-pyruvate lyase (UbiC) -- the first committed enzyme of the ubiquinone biosynthesis from chorismate -- in all analyzed genomes (Figure 2) implies a non-orthologous gene displacement or the existence of an alternative metabolic route to 4-hydroxybenzoate in cyanobacteria. The existence of a novel pathway for the biosynthesis of 4-hydroxybenzoate from tyrosine has been proposed by Booth et al. (Booth, Masri et al. 1960) in 1960 in mammals (Figure 4). Evidence in support of this pathway has recently been obtained in higher plants (Loscher and Heide 1994). However, no genes have been associated with these enzymes in any organism. Genome context analysis in SEED of cyanobacterial and other microbial genomes lacking chorismate-pyruvate lyase (EC 4.1.3.-) may lead to identification of this pathway (important also for the core metabolism of higher eukaryotes) or yet another route to 4-hydroxybenzoate. The absence of the penultimate enzyme of ubiquinone biosynthesis (UbiF) (Figure 2), may constitute another case of a “missing gene” or can be due to alleged pathway modification for plastoquinone production.

References:

1. Booth, A. N., M. S. Masri, et al. (1960). Urinary phenolic acid metabolites of tyrosine. JBC, 235(9): 2649-2652.
2. Dufresne, A., M. Salanoubat, et al. (2003). Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. PNAS, 100(17): 10020-10025.
3. Loscher, R. and L. Heide (1994). Biosynthesis of *p*-hydroxybenzoate from *p*-coumarate and *p*-coumaroyl-CoA in cell-free extracts of *Lithospermum erythrorhizon* cell cultures. Plant Physiol. 106(1): 271-279.
4. Meganathan, R. (2001). Ubiquinone biosynthesis in microorganisms. FEMS Microbiol. Lett. 203(2): 131-139.
6. Olson, R. E. and H. Rudney (1983). Biosynthesis of Ubiquinone. Vit. and Horm., 40: 1-43.

Figure 1. Menaquinone and Phylloquinone biosynthesis in cyanobacteria.

A: Subsystem spreadsheet

B: Subsystem diagram: key intermediates (circles with roman numerals) connected by enzymes (boxes with abbreviations) and reactions (arrows). Presence of genes with respective functions in genome of *Synechocystis* 6803 is shown by green highlighting.

C: Functional roles in subsystem

Basic Spreadsheet											
Organism	Variant Code	menF	menD1	menD2	menC	menE	menB	menH	menA	ubiE	menG
Anabaena variabilis ATCC 29413 [B]	2	195	6185	6185	4532	4531		2730, 3458, 3618, 5930	197, 5476	1687, 4157, 5717, 6230	1232
Crocospaera watsonii WH 8501 [B]	2	1843	4426	4426	5800	5640		2822, 563	1845	59, 6325	988
Gloeobacter violaceus PCC 7421 [B]	4							1217, 2002, 796	1578	3355	127
Nostoc punctiforme [B]	1	6251	4590	4590	6253	6254	4592	3599, 4161, 549, 6812	6252	2472	3427
Nostoc sp. PCC 7120 [B]	1	343	623	623	345	346	2654	2363, 3060, 3992, 422	344	3402	5559
Prochlorococcus marinus MED4 [B]	1	270	1752	1752	273	274	1751	545	272	1957	1937
Prochlorococcus marinus str. MIT 9313 [B]	1	2052	406	406	2050	2049	405	128	2051	1168, 612	275
Prochlorococcus marinus subsp. marinus str. CCMP1375 [B]	2	202	1052	1052	200	199	1051	1787	201		427
Prochlorococcus marinus subsp. pastoris str. CCMP1986 [B]	1	177	604	604	175	174	605	1624	176	413, 918	428
Synechococcus elongatus PCC 7942 [B]	1	1492	764	764	363	364	1972	123, 1813, 1818	696	1710	439
Organism	Variant Code	menF	menD1	menD2	menC	menE	menB	menH	menA	ubiE	menG
Synechococcus sp. WH 8102 [B]	1	2299	992	992	2297	2296	993	676	2298	679	1667
Synechocystis sp. PCC 6803 [B]	1	1550	2397	2397	2280	2353	956	2714, 3250, 561, 943	1451	2700	254
Thermosynechococcus elongatus BP-1 [B]	2	1212	130	130	1173	1220	2456	2065	2195		2372
Trichodesmium erythraeum IMS101 [B]	2	5230	5238	5238	5225	5224		1446, 3557, 5226, 998	5228	762	5107

Intermediates	
I	Chorismate
II	Isochorismate
III	2-Oxoglutarate
IV	Succinate semialdehyde-thiamin diphosphate anion
V	2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate
VI	o-Succinylbenzoate
VII	O-Succinylbenzoyl-CoA
VIII	1,4-dihydroxy-2-naphthoyl-CoA
IX	1,4-dihydroxy-2-naphthoate
X	polyprenylpyrophosphate
XI	2-Demethylmenaquinone
XII	Menaquinone
XIII	Phylloquinone
SAM	S-Adenosyl-L-methionine
SAH	S-Adenosyl-L-homocysteine
TPP	Thiamin pyrophosphate

Abbrev

menF	Menaquinone-specific isochorismate synthase (EC 5.4.4.2)
menD1	2-oxoglutarate decarboxylase (EC 4.1.1.71)
menD2	2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase (EC 4.1.3.-)
menC	O-succinylbenzoate-CoA synthase (EC 4.2.1.-)
menE	O-succinylbenzoic acid--CoA ligase (EC 6.2.1.26)
menB	Naphthoate synthase (EC 4.1.3.36)
menH	Menaquinone biosynthesis related protein, putative DHNA-CoA thioesterase
menA	1,4-dihydroxy-2-naphthoate octaprenyltransferase (EC 2.5.1.-)
ubiE	Ubiquinone/menaquinone biosynthesis methyltransferase UbiE/COQ5 (EC 2.1.1.-)
menG	S-adenosylmethionine:2-demethylmenaquinone methyltransferase (EC 2.1.-.-)

Functional Roles	
Abbrev	Functional Role
UbiC	Chorismate--pyruvate lyase (EC 4.-.-.-)
UbiA	4-hydroxybenzoate polyprenyltransferase (EC 2.5.1.-)
UbiD	3-polyprenyl-4-hydroxybenzoate carboxy-lyase (EC 4.1.1.-)
UbiB	Ubiquinone biosynthesis monooxygenase UbiB
UbiG	Ubiquinone biosynthesis SAM-dependent O-methyltransferase (EC 2.1.1.-)
UbiH	Ubiquinone biosynthesis monooxygenase UbiH/COQ6 (EC 1.14.13.-)
UbiE	Ubiquinone/menaquinone biosynthesis methyltransferase UbiE/COQ8 (EC 2.1.1.-)
UbiF	Ubiquinone biosynthesis monooxygenase UbiF/COQ7 (EC 1.14.13.-)

Figure 2. Subsystem: Ubiquinone Biosynthesis

A: Functional roles in the Subsystem

B: Subsystem spreadsheet: red boxes indicate roles with corresponding genes “missing” from all cyanobacterial genomes

Basic Spreadsheet									
Organism	Variant Code	UbiC	UbiA	UbiD	UbiB	UbiG	UbiH	UbiE	UbiF
Anabaena variabilis ATCC 29413 [B]	3		5850	2498	1243 , 4264 , 590		2007	1687 , 4157 , 5717 , 6230	
Crocospaera watsonii WH 8501 [B]	3		4074		2855 , 4336 , 6350 , 709 , 799	1736		59 , 6325	
Gloeobacter violaceus PCC 7421 [B]	3		3428	3010	1468 , 3124 , 4178 , 4412	4306		3355	
Nostoc punctiforme [B]	3		2178	5176	129 , 1306 , 1847 , 6836	4641 , 624	3556	2472	
Nostoc sp. PCC 7120 [B]	3		3860	1418	4822 , 5267 , 903	323	4581	3402	
Prochlorococcus marinus MED4 [B]	3		1909		1171 , 238 , 458		1477	1957	
Prochlorococcus marinus str. MIT 9313 [B]	3		1322	397	2073 , 2263	1424 , 1468	811	1168 , 612	
Prochlorococcus marinus subsp. marinus str. CCMP1375 [B]	3		455	1045	1880 , 236	1432	904		
Prochlorococcus marinus subsp. pastoris str. CCMP1986 [B]	3		453		1079 , 1710 , 209		831	413 , 918	
Synechococcus elongatus PCC 7942 [B]	3		1273	2568	216		1363	1710	
Organism	Variant Code	UbiC	UbiA	UbiD	UbiB	UbiG	UbiH	UbiE	UbiF
Synechococcus sp. WH 8102 [B]	3		1840	1001	2340 , 2515		811	679	
Synechocystis sp. PCC 6803 [B]	3		2880	298	1093 , 2237 , 2677 , 568	2604 , 333	271	2700	
Thermosynechococcus elongatus BP-1 [B]	3		401	2224	1659 , 2241 , 2284				
Trichodesmium erythraeum IMS101 [B]	3		2897		2608 , 31 , 4203 , 68			762	

Figure 3. Ubiquinone Biosynthesis in *Synechocystis* 6803

Subsystem diagram: key intermediates (circles with roman numerals) connected by enzymes (boxes with abbreviations) and reactions (arrows). Presence of genes with respective functions in the genome of *Synechocystis* 6803 and other organisms is shown by green highlighting. Red boxes indicate enzymes corresponding to “missing genes.” The 4-hydroxybenzoate biosynthesis pathway remains to be identified.

I	Chorismate
II	4-Hydroxybenzoate
III	3-polyprenyl-4-hydroxybenzoate
IV	Polyprenyl diphosphate
V	2-polyprenylphenol
VI	2-polyprenyl-6-hydroxyphenol
VII	2-polyprenyl-6-methoxyphenol
VIII	2-polyprenyl-6-methoxy-1,4-benzoquinone
IX	2-polyprenyl-3-methyl-6-methoxy-1,4-benzoquinone
X	2-polyprenyl-3-methyl-5-hydroxy-6-methoxy-1,4-benzoquinone
XI	Ubiquinole
XII	Ubiquinone
XIII	3,4-dihydroxy-5-polyprenilbenzoate
XIV	3-methoxy-4-hydroxy-5-polyprenilbenzoate
SAM	S-Adenosyl-L-methionine
SAH	S-Adenosyl-L-homocysteine

Abbrev

Abbrev	Functional Role
UbiC	Chorismate--pyruvate lyase (EC 4.-.-.-)
UbiA	4-hydroxybenzoate polyprenyltransferase (EC 2.5.1.-)
UbiD	3-polyprenyl-4-hydroxybenzoate carboxy-lyase (EC 4.1.1.-)
UbiB	Ubiquinone biosynthesis monooxygenase UbiB
UbiG	Ubiquinone biosynthesis SAM-dependent O-methyltransferase (EC 2.1.1.-)
UbiH	Ubiquinone biosynthesis monooxygenase UbiH/COQ6 (EC 1.14.13.-)
UbiE	Ubiquinone/menaquinone biosynthesis methyltransferase UbiE/COQ5 (EC 2.1.1.-)

Figure 4: Proposed pathway for the biosynthesis of 4-hydroxybenzoate from tyrosine: 2, tyrosine; 5, 4-hydroxyphenylpyruvate; 6, 4-hydroxyphenyllactate; 7, 4-hydroxycinnamate or p-coumarate; 8, p-coumaroyl-CoA; 9, 4-hydroxybenzoyl-CoA; 3, hydroxybenzoate (from Meganathan, 2001)

The *in vitro* conversion of tyrosine (Fig. 5, 2) to 4-hydroxyphenylpyruvate (5) and its subsequent reduction to 4-hydroxyphenyllactate (6) has been shown in extracts of rat liver and yeast. Further reactions involved in the conversion of 4-hydroxyphenyllactate (6) to 4-hydroxybenzoate (3) are not known. However, it has been suggested that 4-hydroxyphenyllactate (6) is converted to 4-hydroxycinnamate (7), followed by L-oxidation as the coenzyme A derivative, resulting in the formation of 4-hydroxybenzoate (3) (Olson and Rudney 1983).

Figure 5. Subsystem: Plastoquinone Biosynthesis

A: Subsystem spreadsheet: red boxes indicate roles with corresponding genes “missing” from *Prochlorococcus* and several other cyanobacterial genomes

B: Subsystem diagram

A

Organism	Variant Code	HPPD	HPT	MSBM
Prochlorococcus marinus MED4	3			666
Prochlorococcus marinus str. MIT 9313	3			1779
Prochlorococcus marinus subsp. marinus str. CCMP1375	3			1658
Prochlorococcus marinus subsp. pastoris str. CCMP1986	3			1501
Synechococcus sp. WH 8102	3			2134
Thermosynechococcus elongatus BP-1	3			1725
Gloeobacter violaceus PCC 7421	2	3091	283	
Nostoc punctiforme	2	924	4733	
Organism	Variant Code	HPPD	HPT	MSBM
Anabaena variabilis ATCC 29413	1	6233	5775	1689
Arabidopsis thaliana [E]	1	833	13260, 8892	17735
Crocospaera watsonii WH 8501	1	1498	5070	1145
Nostoc sp. PCC 7120	1	1081	3755	2428
Synechocystis sp. PCC 6803	1	2594	1195	1908
Trichodesmium erythraeum IMS101	1	5369	5002	6205

B

Abbrev	Functional Role
HPPD	4-hydroxyphenylpyruvate dioxygenase (EC 1.13.11.27)
HST	Homogentisate solanyltransferase
MSBM	2-methyl-6-solanyl-1,4-benzoquinone methyltransferase

	Intermediates
I	4-Hydroxyphenylpyruvate
II	Homogentisate
III	Solanyl diphosphate
IV	2-methyl-6-solanyl-1,4-benzoquinone
V	Plastoquinone
SAM	S-Adenosyl-L-methionine
SAH	S-Adenosyl-L-homocysteine

Example #5. Subsystem: Isoprenoid Biosynthesis

Summary

- All cyanobacteria have a functional nonmevalonate pathways of isoprenoid biosynthesis;
- All genes involved (except a few in incomplete genomes) are unambiguously identified based on homology;
- Variations: isomerase IPPDI is absent in all *Prochlorococci* and *Synechococcus* sp. WH 8102. It is optional for DOXP pathway.

Subsystem: Isoprenoid Biosynthesis

MEVALONATE PATHWAY	
I	Acetoacetyl-CoA
II	Hydroxymethylglutaryl-CoA
III	Mevalonate
IV	Phosphomevalonate
V	Diphosphomevalonate
NONMEVALONATE PATHWAY	
VI	1-Deoxy-D-xylulose 5-phosphate
VII	2C-methyl-D-erythritol 4-phosphate
VIII	4-Diphosphocytidyl-2C-methyl-D-erythritol
IX	4-Diphosphocytidyl-2C-methyl-D-erythritol 2-
X	2C-Methyl-D-erythritol 2,4-cyclodiphosphate
XI	1-Hydroxy-2-methyl-2(E)-butenyl 4-diphosphate
COMMON COMPOUNDS	
IPP	Isopentenyl Pyrophosphate
DMAPP	Dimethylallyl Pyrophosphate
OTHER COMPOUNDS	
AcCoA	Acetyl-CoA
Pyruvate	
GA3P	D-glyceraldehyde-3-phosphate

Mevalonate pathway	
ACAT	Acetyl-CoA acetyltransferase (EC 2.3.1.9)
HMGCS	Hydroxymethylglutaryl-CoA synthase (EC 2.3.3.10)
HMGCR	Hydroxymethylglutaryl-CoA reductase (EC 1.1.1.34)
MEVK	Mevalonate kinase (EC 2.7.1.36)
PMEVK	Phosphomevalonate kinase (EC 2.7.4.2)
DPMVD	Diphosphomevalonate decarboxylase (EC 4.1.1.33)
DOXP pathway	
DXS	1-deoxy-D-xylulose 5-phosphate synthase (EC 2.2.1.7)
ISPC	1-deoxy-D-xylulose 5-phosphate reductoisomerase (EC 1.1.1.267)
ISPD	2-C-methyl-D-erythritol 4-phosphate cytidyltransferase (EC 2.7.7.60)
ISPE	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (EC 2.7.1.148)
ISPF	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (EC 4.6.1.12)
ISPG	1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase (EC 1.17.4.3)
ISPH	4-hydroxy-3-methylbut-2-enyl diphosphate reductase (EC 1.17.1.2)
Common enzymes	
IPDDI	Isopentenyl-diphosphate delta-isomerase (EC 5.3.3.2)
GTT	Geranyltranstransferase (EC 2.5.1.10)

Cyanobacteria:
DOXP pathway

Subsystem: Isoprenoid Biosynthesis
Organism: Synechocystis sp

Genome ID	Organism	Variant Code	ACAT	HMGCS	HMGCR	MEVK	PMEVK	DPMVD	IPDDI	DXS	ISPC	ISPD	ISPE	ISPF	ISPG	ISPH	GTT
1140.1	Synechococcus elongatus PCC 7942 [B]	28							1244	?	837	?	2232	504	1865	2445	125
84588.1	Synechococcus sp. WH 8102 [B]	29							1287	693	1842	1048	1605	1169	249	737	
1148.1	Synechocystis sp. PCC 6803 [B]	28	1303						1802	510	2214	1838	86	887	1138	2190	111, 3167
197221.1	Thermosynechococcus elongatus BP-1 [B]	2							1402	621	1040	604	499	2034	?	1041	20
63737.1	Nostoc punctiforme [B]	2		4143					2273	4226	1096	1250	577	4479	1897	2962	5170, 745
103690.1	Nostoc sp. PCC 7120 [B]	29							4898	910	4658	5474	3537	4190	?	1295	524
74547.1	Prochlorococcus marinus str. MIT 9313 [B]	28								684	1158	1324	620	356	776	1848	1107
167539.1	Prochlorococcus marinus subsp. marinus str. CCMP1375 [B]	28								926	1234	453	762	1351	1013	296	1127
59919.1	Prochlorococcus marinus subsp. pastoris str. CCMP1986 [B]	28								904	1138	451	929	1276	673	262	1066

Example #6. Subsystem: Lysine Biosynthesis DAP Pathway

Summary

- Three genes are still uncharacterized (missing) in all cyanobacteria. In the absence of this information it is impossible to choose between possible alternatives: succinylation vs acetylation pathways
- **Candidate genes predicted by homology searches and occurrence profile analysis argue in favor of acetylation pathway**
- Experimental verification is suggested

Subsystem: Lysine Biosynthesis DAP Pathway

Subsystem: Lysine Biosynthesis DAP Pathway

Functional Roles

Column	Abbrev	Functional Role
1	AK	Aspartokinase (EC 2.7.2.4)
2	ASD	Aspartate-semialdehyde dehydrogenase (EC 1.2.1.11)
3	DHPS	Dihydrodipicolinate synthase (EC 4.2.1.52)
4	DHPR	Dihydrodipicolinate reductase (EC 1.3.1.26)
5	THPST	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase (EC 2.3.1.117)
6	SDAPA	N-succinyl-L,L-diaminopimelate aminotransferase (EC 2.6.1.17)
7	DAPDS	N-succinyl-L,L-diaminopimelate desuccinylase (EC 3.5.1.18)
8	THPAT	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-acetyltransferase (EC 2.3.1.89)
9	ADAPA	N-acetyl-L,L-diaminopimelate aminotransferase (EC 2.6.1.-)
10	DAPDA	N-acetyl-L,L-diaminopimelate deacetylase (EC 3.5.1.18)
11	DAPDH	Meso-diaminopimelate D-dehydrogenase (EC 1.4.1.16)
12	DAPE	Diaminopimelate epimerase (EC 5.1.1.7)
13	DAPE1	Diaminopimelate epimerase alternative form predicted for S.aureus (EC 5.1.1.7)
14	DAPDC	Diaminopimelate decarboxylase (EC 4.1.1.20)
15	ArgD	N-succinyl-L,L-diaminopimelate aminotransferase alternative (EC 2.6.1.17)
16		
17		
18		
19		
20		

Subsets of Roles

Subset	Includes These Roles
*DAPE	12,13
*SDAPAT	6,15
Acetylation_pathway	1,2,3,4,8,9,10,12,14
DAP_dehydrogenase_pathway	1,2,3,4,11,14
Succinylation_pathway	1,2,3,4,5,6,7,12,14

Example #1: Succinylation pathway

Subsystem: Lysine Biosynthesis DAP Pathway

Organism: *E.coli*

Genome ID	Organism	Variant Code	AK	ASD	DHPS	DHPR	THPST	*SDAPAT	DAPDS	THPAT	ADAPAT	DAPDA	DAPDH	*DAPE	DAPDC
83333.1	Escherichia coli K12 [B]	1	2,3859,3933	2293,3367	2447	31	166	2352-6,3294-15	2440				3735-12		

Example #2: Succinylation and DAPDH pathways

Subsystem: Lysine Biosynthesis DAP Pathway

Organism: *C.glutamicum*

Genome ID	Organism	Variant Code	AK	ASD	DHPS	DHPR	THPAT	*SDAPAT	DAPDS	THPAT	ADAPAT	DAPDA	DAPDH	*DAPE	DAPDC
96627.1	Corynebacterium glutamicum ATCC 13032 [B]	3	247	248	1896	1898	1061 , 1063	1058-6	1064				2528	1868-12	1133

Example #3: Acetylation pathway

Subsystem: Lysine Biosynthesis DAP Pathway
Organism: *B.subtilis*

Genome ID	Organism	Variant Code	AK	ASD	DHPS	DHPR	THPAT	*SDAPAT	DAPDS	THPAT	ADAPAT	DAPDA	DAPDH	*DAPE	DAPDC
224308.1	Bacillus subtilis subsp. subtilis str. 168 [B]	2	1679, 2850, 380	1678	1680	2253				1420	1402	1421		3223-12	2343

Subsystem: Lysine Biosynthesis DAP Pathway

Missing genes in cyanobacteria

165597.1	Crocospaera watsonii WH 8501 [B]	0	5778	3127	3126	1479		1945-15						616-12	1537
Genome ID	Organism	Variant Code	AK	ASD	DHPS	DHPR	THPST	*SDAPAT	DAPDS	THPAT	ADAPAT	DAPDA	DAPDH	*DAPE	DAPDC
251221.1	Gloeobacter violaceus PCC 7421 [B]	0	1774	1017	1018	1206		4108-6, 547-15						815-12	110
1140.1	Synechococcus elongatus PCC 7942 [B]	0	344	1162	1161	1444		287-15				2280		1238-12	
84588.1	Synechococcus sp. WH 8102 [B]	0	67	63	64	814		1629-15						1249-12	931
1148.1	Synechocystis sp. PCC 6803 [B]	0	2965	2899	2900	65		580-15						3074-12	2861
240292.1	Anabaena variabilis ATCC 29413 [B]	0	5977	938	939	3048		4903-6, 2474-15						2740-12, 5034-12	4982, 69
103690.1	Nostoc sp. PCC 7120 [B]	0	3951	3987	3986	2849		3072-6, 1390-15						2355-12, 5148-12	3304
74547.1	Prochlorococcus marinus str. MIT 9313 [B]	0	73	64	65	813		331-15						715-12	1057

Subsystem: Lysine Biosynthesis DAP Pathway
Inferred acetylation variant

Example #7. Subsystem: Coenzyme A Biosynthesis

Summary

- All enzymes, except for Aspartate 1-decarboxylase (ASPDC), of the complete *de novo* biosynthesis of CoA are identified in all cyanobacteria. The gene for ASPDC is missing in *Prochlorococci*, *Synechococcus* sp. WH 8102, *Thermosynechococcus elongatus*.
- This may suggest an alternative form of ASPDC (to be identified) or an alternative source of beta-alanine (more likely) in these organisms
- Functional prediction: all cyanobacteria lack a “classic” form of pantothenate kinase EC 2.7.1.33 (coaA). Based on recent findings in *B. subtilis* we propose that this enzyme is encoded in *Synechocystis* 6803 by the previously uncharacterized gene **slr0812**, and by its orthologs in all other cyanobacterial genomes. It is annotated in SEED as Pantothenate kinase type III, CoaX-like (EC 2.7.1.33), while in all other genomic databases it appears as: hypothetical protein slr0812, putative transcriptional regulators, or homologs of Bvg accessory factor.

Subsystem: Coenzyme A Biosynthesis

Subsystem: Coenzyme A Biosynthesis

of Genomes present: 266

Functional Roles

Column	Abbrev	Functional Role
1	ASPDC	Aspartate 1-decarboxylase (EC 4.1.1.11)
2	KPHMT	3-methyl-2-oxobutanoate hydroxymethyltransferase (EC 2.1.2.11)
3	KPRED	2-dehydropantoate 2-reductase (EC 1.1.1.169)
4	KARED	Ketol-acid reductoisomerase (EC 1.1.1.86)
5	PBAL	Pantoate--beta-alanine ligase (EC 6.3.2.1)
6	PANF	Pantothenate:Na ⁺ symporter (TC 2.A.21.1.1)
7	PANK	Pantothenate kinase (EC 2.7.1.33)
8	PANK2	Pantothenate kinase type II, eukaryotic (EC 2.7.1.33)
9	PANK3	Pantothenate kinase type III, CoaX-like (EC 2.7.1.33)
10	PPCS	Phosphopantothenoylcysteine synthetase (EC 6.3.2.5)
11	PPCDC	Phosphopantothenoylcysteine decarboxylase (EC 4.1.1.36)
12	PPAT	Phosphopantetheine adenylyltransferase (EC 2.7.7.3)
13	PPAT2	Phosphopantetheine adenylyltransferase, type II eukaryotic (EC 2.7.7.3)
14	DPCK	Dephospho-CoA kinase (EC 2.7.1.24)
15		

Subsystem: Coenzyme A Biosynthesis

organism: *Homo sapiens*

Functional variant 2

Subsystem: Coenzyme A Biosynthesis cyanobacteria

Predicted: hypothetical protein) slr0812

Basic Spreadsheet																
Genome ID	Organism	Variant Code	ASPDC	KPHMT	KPRED	KARED	PBAL	PANF	PANK	PANK2	PANK3	PPCS	PPCDC	PPAT	PPAT2	DPCK
251221.1	Gloeobacter violaceus PCC 7421 [B]	3*	2845 , 4228	762		2657	1806	3452			2994	1548	1548	847		2134
1140.1	Synechococcus elongatus PCC 7942 [B]	34*	?	1679		870	487				1123			1316		2264
84588.1	Synechococcus sp. WH 8102 [B]	13*	?	1640		1645	2020	2446 , 614			2156	298	298	1243		2345
1148.1	Synechocystis sp. PCC 6803 [B]	13	?	2843		2955	1586	1767			1541	1365	1365	2568		2903
197221.1	Thermosynechococcus elongatus BP-1 [B]	31*	?	279		2253	2449				1148	2465	2465	1164		897
63737.1	Nostoc punctiforme [B]	3*	2073	1008	1922	6429	4100	2130			6828	1669	1669	4607		4738
103690.1	Nostoc sp. PCC 7120 [B]	3*	3876	1829	1629	2622	3243	1985			4203	3418	3418	5007		2061
74547.1	Prochlorococcus marinus str. MIT 9313 [B]	31*	?	318		313	1683	1496			1617	1796	1796	721		2127
167539.1	Prochlorococcus marinus subsp. marinus str. CCMP1375 [B]	31*	?	1381		1386	1743	1632			94	258	258	950		53
59919.1	Prochlorococcus marinus subsp. pastoris str. CCMP1986 [B]	31*	?	1306		1311	1586				80	228	228	881		49
Genome ID	Organism	Variant Code	ASPDC	KPHMT	KPRED	KARED	PBAL	PANF	PANK	PANK2	PANK3	PPCS	PPCDC	PPAT	PPAT2	DPCK
167540.1	Prochlorococcus marinus MED4 [B]	31*	?	916		910	584				372	219	219	1414		406

Subsystem: Coenzyme A Biosynthesis

organism: *Synechocystis sp.*

Example #8. Subsystem: FMN/FAD Biosynthesis

Summary

- Complete pathway is present in all cyanobacteria
- Pyrimidine phosphatase remains a missing gene.
- Notably, not a single gene of riboflavin biosynthesis is clustered in any cyanobacteria, while in the majority of other organisms these genes occur in conserved chromosomal clusters.

Subsystem: FMN and FAD biosynthesis

Functional Roles

Column	Abbrev	Functional Role
1	GTPCH2	GTP cyclohydrolase II (EC 3.5.4.25)
2	Pyr_D	Diaminohydroxyphosphoribosylaminopyrimidine deaminase (EC 3.5.4.26)
3	Pyr_De	Pyrimidine deaminase eukaryotic (EC 3.5.4.26)
4	Pyr_Da	Pyrimidine deaminase archaeal predicted (EC 3.5.4.26)
5	Pyr_R	5-amino-6-(5-phosphoribosylamino)uracil reductase (EC 1.1.1.193)
6	Pyr_P	5-Amino-6-(5'-phosphoribitylamino)uracil phosphatase
7	DHBPS	3,4-dihydroxy-2-butanone 4-phosphate synthase
8	DMRS	6,7-dimethyl-8-ribityllumazine synthase (EC 2.5.1.9)
9	RSA	Riboflavin synthase alpha chain (EC 2.5.1.9)
10	RSAA	Riboflavin synthase archaeal (EC 2.5.1.9)
11	YpaA	Riboflavin transporter YpaA
12	RK	Riboflavin kinase (EC 2.7.1.26)
13	FMNAT	FMN adenylyltransferase (EC 2.7.7.2)
14	FMNAT2	FMN adenylyltransferase, type 2 eukaryotic (EC 2.7.7.2)
15	GTPCH3	GTP cyclohydrolase III archaeal hypothesized
16	NTPPP	Nucleotide triphosphate pyrophosphatase (EC 3.6.1.19)

Subsets of Roles

Subset	Includes These Roles
*FMNAT	13,14
*Pyr_D	2,3,4
*RSA	9,10
Archaeal_branch	15,16

Organism: *E.coli*

***NB: no transport of exogenous ribofavin in wt strain;
All genes are essential (see Gerdes et al., 2003)***

GTP	Guanosine triphosphate
I	2,5-Diamino-6-ribosylamino-4(3H)-pyrimidinone 5'-phosphate
II	5-Amino-6-ribosylamino-2,4(1H,3H)-pyrimidinedione 5'-phosphate
III	5-Amino-6-ribitylamino-2,4(1H,3H)-pyrimidinedione 5'-phosphate
IV	5-Amino-6-ribitylamino-2,4(1H,3H)-pyrimidinedione
V	6,7-Dimethyl-8-ribityl-lumazine
VI	2,5-diamino-6-ribosylamino-4(3H)-pyrimidinone 5-triphosphate
VI	Ribulose 5-phosphate
VII	L-3,4-Dihydroxy-2-butanone 4-phosphate
B2	Riboflavin
FMN	Flavin Mononucleotide
FAD	Flavin Adenine Dinucleotide
VIII	2,5-diamino-6-ribitylamino-4(3H)-pyrimidinone 5-phosphate
IX	2,5-diamino-6-ribosylamino-4(3H)-pyrimidinone 5-triphosphate

Organism	Variant Code	GTPCH2	*Pyr_D	Pyr_R	Pyr_P	DHBPS	DMRS	*RSA	YpaA	RK	*FMNAT
Escherichia coli E2348 [B]	1*	1909	3189-2	3189		5094	3188	438-9		3556	3556-13
Escherichia coli K12 [B]	1	1264	411-2	411		2989	412	1648-9		25	25-13

Subsystem: FMN/FAD Biosynthesis

Organism: *B. subtilis*

NB: *ypaA* was predicted and confirmed as riboflavin transporter;

Only last two genes are essential (see Kobayashi et al., 2003).

Same pattern in *S. aureus*, *S. pneumoniae* and some others...

GTP	Guanosine triphosphate
I	2,5-Diamino-6-ribosylamino-4(3H)-pyrimidinone 5'-phosphate
II	5-Amino-6-ribosylamino-2,4(1H,3H)-pyrimidinedione 5'-phosphate
III	5-Amino-6-ribitylamino-2,4(1H,3H)-pyrimidinedione 5'-phosphate
IV	5-Amino-6-ribitylamino-2,4(1H,3H)-pyrimidinedione
V	6,7-Dimethyl-8-ribityl-lumazine
VI	2,5-diamino-6-ribosylamino-4(3H)-pyrimidinone 5-triphosphate
VII	Ribulose 5-phosphate
B2	Riboflavin
FMN	Flavin Mononucleotide
FAD	Flavin Adenine Dinucleotide
VIII	2,5-diamino-6-ribitylamino-4(3H)-pyrimidinone 5-phosphate
IX	2,5-diamino-6-ribosylamino-4(3H)-pyrimidinone 5-triphosphate

Organism	Variant Code	GTPCH2	*Pyr_D	Pyr_R	Pyr_P	DHBPS	DMRS	*RSA	YpaA	RK	*FMNAT
<i>Staphylococcus epidermidis</i> ATCC 12228 [B]	11	1439	1441-2	1441		1439	1438	1440-9	1174	949	949-13
<i>Clostridium acetobutylicum</i> ATCC 824 [B]	11*	759	757-2	757		759	760	758-9	2974	1960	1960-13
<i>Clostridium botulinum</i> ATCC 3502 [B]	11*	695	697-2	697		695, 947	694	696-9	20	316	316-13

Subsystem: FMN/FAD Biosynthesis

Organism: *Synechocystis* sp.

NB: a completely straightforward case, similar to B.subtilis, including ypaA

Organism	Variant Code	GTPCH2	*Pyr_D	Pyr_R	Pyr_P	DHBPS	DMRS	*RSA	YpaA	RK	*FMNAT
Synechocystis sp. PCC 6803 [B]	1	1611	2326-2	2326		1611	988	2058-9		1117	1117-13

Subsystem: FMN/FAD Biosynthesis
cyanobacteria

Genome ID	Organism	Variant Code	GTPCH2	*Pyr_D	Pyr_R	Pyr_P	DHBPS	DMRS	*RSA	YpaA	RK	*FMNAT
251221.1	Gloeobacter violaceus PCC 7421 [B]	1	988	1250-2	1250		988	1042	3604-9		217	217-13
1140.1	Synechococcus elongatus PCC 7942 [B]	1	761	2331-2	2331		761	1544	945-9		2065	2065-13
84588.1	Synechococcus sp. WH 8102 [B]	1	1258	1585-2	1585		1258	79	1858-9		1598	1598-13
1148.1	Synechocystis sp. PCC 6803 [B]	1	1611	2326-2	2326		1611	988	2058-9		1117	1117-13
197221.1	Thermosynechococcus elongatus BP-1 [B]	1	1726	2151-2	2151		1726	67	1685-9		1555	1555-13
63737.1	Nostoc punctiforme [B]	1	2652	5380-2	5380		2652	134	4049-9		5488	5488-13
103690.1	Nostoc sp. PCC 7120 [B]	1	3843	393-2	393		3843	4300	5565-9		5155	5155-13
74547.1	Prochlorococcus marinus str. MIT 9313 [B]	1	707	375-2	375		707	81	1340-9		364	364-13
167539.1	Prochlorococcus marinus subsp. marinus str. CCMP1375 [B]	1	941	1336-2	1336		941	1800	438-9		1344	1344-13
59919.1	Prochlorococcus marinus subsp. pastoris str. CCMP1986 [B]	1	890	1261-2	1261		890	1639	439-9		1269	1269-13
Genome ID	Organism	Variant Code	GTPCH2	*Pyr_D	Pyr_R	Pyr_P	DHBPS	DMRS	*RSA	YpaA	RK	*FMNAT
167540.1	Prochlorococcus marinus MED4 [B]	1	1405	960-2	960		1405	532	1924-9		953	953-13

Example #9. Subsystem: Chlorophyll Biosynthesis

Summary

- Complete pathway is present in all cyanobacteria
- 4-vinyl-protochlorophyllide reductase (EC 1.3.1.75) remains a missing gene in cyanobacteria and higher plants. The gene *bchJ* encoding this enzyme in *Rhodobacter* sp. has no orthologs in other groups of chlorophyll-containing organisms.
- A candidate gene for the missing role of Chlorophyll/chlorophyllide a oxygenase (EC 1.14.13.-) has been identified in *Prochlorococci* based on long-range similarity and clustering on the chromosome (PMT0774 in *Prochlorococcus marinus* str. MIT 9313)

Subsystem: Chlorophyll Biosynthesis -- Plants

PMgCD	Protoporphyrin IX Mg-chelatase subunit D (EC 6.6.1.1)
PMgCH	Protoporphyrin IX Mg-chelatase subunit H (EC 6.6.1.1)
PMgCI	Protoporphyrin IX Mg-chelatase subunit I (EC 6.6.1.1)
PMgMT	Mg-protoporphyrin O-methyltransferase (EC 2.1.1.11)
ChIEAn	Mg-protoporphyrin IX monomethyl ester oxidative cyclase
ChIEAe	Mg protoporphyrin IX monomethyl ester oxidative cyclase
4VPCR	4-vinyl-protochlorophyllide reductase (EC 1.3.1.75)
ChlL	Light-independent protochlorophyllide reductase iron-sulfur
ChlN	Light-independent protochlorophyllide reductase subunit
ChlB	Light-independent protochlorophyllide reductase subunit
POR	Light-dependent protochlorophyllide reductase (EC 1.3.1.1)
ChlG	Chlorophyll a synthase ChlG (EC 2.5.1.62)
CAO	Chlorophyll a, chlorophyllide a oxygenase (EC 1.14.13.-)
HOX	Heme oxygenase (EC 1.14.99.3)

A. Committed (connecting) intermediates

I	Protoporphyrin IX
II	Mg-Protoporphyrin IX
III	Mg-Protoporphyrin IX 6-monomethyl ester
IV	Divinyl Protochlorophyllide
V	Protochlorophyllide
VI	Divinyl Chlorophyllide a
VII	Chlorophyllide a
VIII	Chlorophyll a
IX	Chlorophyll b
X	Protoheme (heme)
XI	biliverdin IX alpha

Porphyrin, Heme,
Siroheme Biosynthesis

Subsystem: Chlorophyll Biosynthesis -- Cyanobacteria (Synech 6803)

PMgCD	Protoporphyrin IX Mg-chelatase subunit D (EC 6.6.1.1)
PMgCH	Protoporphyrin IX Mg-chelatase subunit H (EC 6.6.1.1)
PMgCI	Protoporphyrin IX Mg-chelatase subunit I (EC 6.6.1.1)
PMgMT	Mg-protoporphyrin O-methyltransferase (EC 2.1.1.11)
ChIEAn	Mg-protoporphyrin IX monomethyl ester oxidative cyclase
ChIEAe	Mg protoporphyrin IX monomethyl ester oxidative cyclase
4VPCR	4-vinyl-protochlorophyllide reductase (EC 1.3.1.75)
ChlL	Light-independent protochlorophyllide reductase iron-sulfur
ChlN	Light-independent protochlorophyllide reductase subunit
ChlB	Light-independent protochlorophyllide reductase subunit
POR	Light-dependent protochlorophyllide reductase (EC 1.3.1.75)
ChlG	Chlorophyll a synthase ChlG (EC 2.5.1.62)
CAO	Chlorophyll a, chlorophyllide a oxygenase (EC 1.14.13.-)
HOX	Heme oxygenase (EC 1.14.99.3)

A. Committed (connecting) intermediates

I	Protoporphyrin IX
II	Mg-Protoporphyrin IX
III	Mg-Protoporphyrin IX 6-monomethyl ester
IV	Divinyl Protochlorophyllide
V	Protochlorophyllide
VI	Divinyl Chlorophyllide a
VII	Chlorophyllide a
VIII	Chlorophyll a
IX	Chlorophyll b
X	Protoheme (heme)
XI	biliverdin IX alpha