|  |  |
| --- | --- |
| |  | | --- | |  |   lAWRENCE bERKELEY NATIONAL lABORATORIES |
| *Microcoleus vaginatus PCC 9802* |
| Annotation Reports |
|  |
|  |
|  |

Contents

[Amino Acid Biosynthesis 2](#_Toc365218234)

[Histidine 2](#_Toc365218235)

[Alanine 14](#_Toc365218236)

[Glutamine & Glutamate 20](#_Toc365218237)

[Arginine 29](#_Toc365218238)

[Asparagine 37](#_Toc365218239)

[Glycine 44](#_Toc365218240)

[Leucine, Valine and Isoleucine 51](#_Toc365218241)

[Lysine, Threonine amd Methionine 64](#_Toc365218242)

[Phenylalanine, Tyrosine and Tryptophan 82](#_Toc365218243)

[Proline 102](#_Toc365218244)

[Serine 108](#_Toc365218245)

[Fermentation 119](#_Toc365218246)

[Bifidobacterium Shunt 119](#_Toc365218247)

[Heterolactic Acid Fermentation 133](#_Toc365218248)

[Homolactic Fermentation 150](#_Toc365218249)

[Mixed Acid Fermentation 163](#_Toc365218250)

[Photosynthesis 179](#_Toc365218251)

[Antennae Protein Photosynthesis Complex 179](#_Toc365218252)

[Light Reactions 191](#_Toc365218253)

[Calvin Cycle 221](#_Toc365218254)

[Nitrogen Metabolism 235](#_Toc365218255)

[Ammonium Assimilation 235](#_Toc365218256)

[Nitrate Assimilation 244](#_Toc365218257)

[Nitrogen Fixation 253](#_Toc365218258)

[Glycogen Biosynthesis 256](#_Toc365218259)

[Glycolysis 264](#_Toc365218260)

[Hydrogenase 276](#_Toc365218261)

[Pentose Phosphate Pathway 278](#_Toc365218262)

[TCA Cycle 287](#_Toc365218263)

[Sulfur and Sulfate Reduction 306](#_Toc365218264)

[Peptidoglycan Biosynthesis 310](#_Toc365218265)

# Amino Acid Biosynthesis

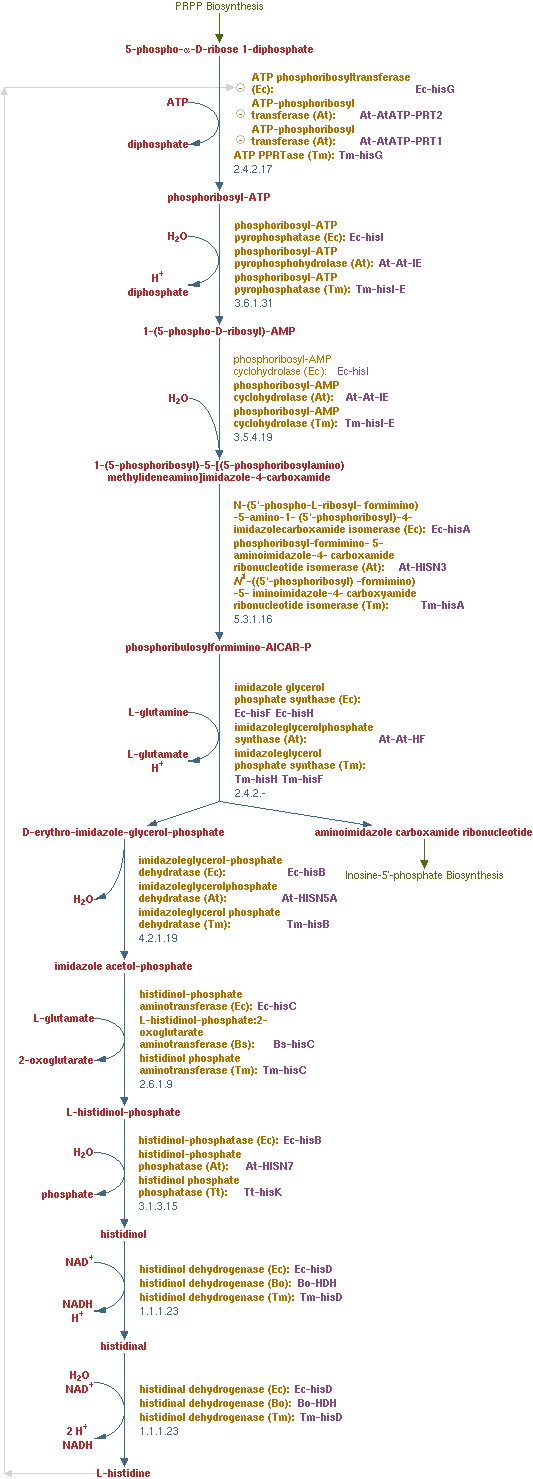
## Histidine

Histidine is an essential amino acid in humans and many other eukaryotes, and such cannot be synthesized by those organisms. Some other organisms however, are able to synthesize their own histidine. *Microcoleus vaginatus* is such an organism that is predicted to be able to synthesize its own histidine.

The genes necessary for histidine biosynthesis have been identified in many bacteria, fungi, plants, and archaea. The pathway in all of these organisms is identical, with small differences in some of the enzymes used. The main differences are found with the enzyme histidinol-phosphate phosphatase (EC 3.1.3.15). Some organisms, including *Microcoleus vaginatus*, have a bifunctional enzyme, with the C-terminal domain exhibiting imidazoleglycerol-phosphate dehydratase (EC 4.2.1.19) activity. These enzymes belong to the DDDD phosphohydrolase/phosphotransferase superfamily. Other organisms, on the other hand, have separate monofunctional enzymes catalyzing these two reactions. Monofunctional histidinol-phosphate phosphatases belong to either the PHP (polymerase and histidinol phosphatase) superfamily, or to the DDDD superfamily.

In all organisms operating the pathway, the enzyme catalyzing the first step is feedback-inhibited by L-histidine so that the organism does not synthesize too much of it.

**Automated Annotation**



S Fig. 1: MetaCyc Pathway Map of Histidine Biosynthesis. Higher quality picture at http://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=HISTSYN-PWY

**ATP phosphoribosyltransferase**

There were two genes predicted to code for this protein. The first was predicted by IMG/ER, while the second was found as a second best hit from a reverse BLAST search.

Gene OID: 2505165694 EC: 2.4.2.17

This gene is predicted to code for ATP phosphoribosyltransferase. A NCBI BLAST search gave only other ATP phosphoribosyltransferases as hits [[9]](http://198.128.28.118/genome/converter.php#ref-9) with e-values from 5E-119 to 1E-38 and 38% to 76% identity. A PDB BLAST search gave as a top hit an article on ATP phosphoribosyltransferase structure [[3]](http://198.128.28.118/genome/converter.php#ref-3) with an e-value of 3.01748E-22 and 31% identity. Out of the 19 active sites mentioned in the article, 9 of them matched. The 10 mismatches were I47L, F48L, K50R, N52H, F57Y, I62Q, D77K, Y82A, E138Y and V147T. Some of these mismatches could be explained because the residues are similar, such as I47L, K50R, and F57Y. For the other mismatches, the article was searched to see if it mentioned that residue. The F48, D77 and Y82 residues were not mentioned as having specific catalytic function. N52 was mentioned for its hydrogen bonding, which H52 can participate in as well. I62 was mentioned as important for its hydrophobicity, but Q62 is hydrophillic. E138 is mentioned as important for its hydrogen bonding, and Y138 can hydrogen bond too. V147 was mentioned for its hydrophobicity, but T147 is hydrophillic. For the residues not mentioned and the residues replaced with dissimilar amino acids, a multiple sequence alignment was done on T-Coffee with reviewed cyanobacterial FASTA sequences from UniProt.

This alignment showed that at site 48, the L48 residue was completely conserved among other cyanobacteria. At site 77, no alignment could be forced between the PDB FASTA sequence and the other cyanobacterial FASTA sequences. At site 82, all but one cyanobacteria had the A82 residue. At site 62, the Q62 residue was completely conserved among other cyanobacteria. Finally, at site 147, the T147 residue is completely conserved among other cyanobacteria.

A reverse BLAST gave this gene as the top hit. From this result and the other data, it is concluded that the gene codes for the enzyme.

Gene OID: 2505166516 EC: 2.4.2.17

This gene is predicted to code for ATP phosphoribosyltransferase regulatory subunit. A NCBI BLAST search gave ATP phosphoribosyltransferase regulatory subunits as top hits {{NCBI: Q8YT12}} with e-values ranging from 0 to 2E-26, with percent identities from 29% to 65% identity. A PDB BLAST search gave as the top hit an unpublished article on ATP phosphoribosyltransferase (PDB: 3OD1) with an e-value of 7.13299E-28 and 31% identity. From these considerations, the gene is predicted to code for ATP phosphoribosyltransferase regulatory subunit.

**Phosphoribosyl-ATP pyrophosphohydrolase**

Gene OID: 2505167005 EC: 3.6.1.31

This gene is expected to code for phosphoribosyl-ATP-pyrophosphohydrolase. A NCBI BLAST search gave hits to histidine biosynthesis bifunctional proteins {{NCBI: Q8YS28}} with e-values from 2E-121 to 2E-35 and 39% to 76% identity. A PDB BLAST search gave as a top hit an article on phosphoribosyl-ATP-pyrophosphohydrolase crystal structure [[8]](http://198.128.28.118/genome/converter.php#ref-8) with an e-value of 3.08378E-6 and 35% identity. Out of the active sites mentioned in the article, 7 out of 9 of them matched. The 2 mismatches were Q70D and E74H. The article mentioned that site 70 is an aspartate in all organisms save actinobacteria, so it makes sense that *Microcoleus vaginatus*, a cyanobacteria, has the D70 residue. The E74 was not mentioned in the article to have specific catalytic function. In addition, a reverse BLAST search gave this gene as the top hit, so accordingly the gene is predicted to code for phosphoribosyl ATP pyrophosphohydrolase.

**Phosphoribosyl-AMP cyclohydrolase**

Gene OID: 2505167005 EC: 3.5.4.19

This gene is also expected to code for phosphoribosyl-AMP cyclohydrolase. A NCBI BLAST search gave the same hits to histidine biosynthesis bifunctional proteins {{NCBI: Q8YS28}}. These protesins are bifunctional for phosphoribosyl-ATP-pyrophosphohydrolase and phophoribosyl-AMP cyclohydrolase function. A PDB BLAST search gave as the top hit an article on phosphoribosyl-AMP cyclohydrolase crystal structure [[4]](http://198.128.28.118/genome/converter.php#ref-4) with an e-value of 4.29281E-24 and 54% identity. Out of the 14 active sites mentioned in the article, 13 of them matched. The sole mismatch was T61R. T61 is only mentioned in the article as contributing a hydrogen bond with its main-chain nitrogen, which isn’t specific to T61. Also, a reverse BLAST search gave this gene as the top hit, so this gene is predicted to code for phosphoribosyl-AMP cyclohydrolase.

**Phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase**

Gene OID: 2505169282 EC: 5.3.1.16

This gene is predicted to code for the protein. The NCBI BLAST search gave only

1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino] imidazole-4-carboxamide  
isomerases as hits [[13]](http://198.128.28.118/genome/converter.php#ref-13) with e-values ranging from 2E-139 to 9E-75 and 49% to 73% identity. A PDB BLAST search gave an article on this protein as a top hit [[5]](http://198.128.28.118/genome/converter.php#ref-5) with an e-value of 8.09044E-39 and 39% identity. Out of the active sites mentioned in the article, 10 out of 14 matched. The mismatches were S81G, R139K, R141A and D148S. In the article, S81 was mentioned as important for binding the substrate via hydrgon bonding, which G81 cannot do. R139, R141 and D148 are not mentioned as having specific catalytic function. These findings are inconclusive, so a multiple sequence alignment was done on T-Coffee, using reviewed cyanobacterial FASTA sequences from UniProt.

The alignment shed some light on residues 81, 139, 141 and 148. At site 81, the G81 residue was completely conserved among the other cyanobacteria. At site 139, there were cyanobacteria with the R139 residue, but there were more with the K139 residue. At site 141, no other cyanobacteria had the R141 bacteria, but most of them had the A141 residue. Finally, at site 148, no other cyanobacteria had the D148 residue, but most had the S148 residue.

From these results and the fact that a reverse BLAST hit gives this gene as a top hit, this gene is predicted to code for the enzyme.

**Imidazole glycerol phosphate synthase**

EC: 2.4.2.-

This protein has two subunits, subunit H and subunit F. IMG/ER predicted a gene to code for subunit H, while the gene coding for subunit F was found from a reverse BLAST search. Both subunits are predicted to have genes coding for them.

Subunit HisH

Gene OID: 2505168714

This gene is predicted to code for subunit HisH. A NCBI BLAST search gave only imidazole glycerol phosphate synthases (subunit HisH) as hits {{NCBI: B0JFQ9}} with e-values ranging from 1E-122 to 2E-47 and and percent identities ranging from 42% to 78%. A PDB BLAST search gave an article on the crystal structure of imidazole glycerol phosphate synthase as a top hit [[12]](http://198.128.28.118/genome/converter.php#ref-12) with an e-value of 9.28457E-32 and 40% identity. 8 out of the 8 active sites mentioned in the article matched, and a reverse BLAST search gave this gene as the top hit, so this gene is predicted to code for subunit HisH.

Subunit HisF

Gene OID: 2505166399

This gene is predicted to code for subunit HisF. A NCBI BLAST search gave only imidazole glycerol phosphate synthases (subunit HisF) as hits {{NCBI: Q8YT31}} with e-values from 5E-155 to 8E-105 and 58% to 86% identity. A PDB BLAST search gave an article on imidazole glycerol phosphate synthase subunit HisF crystal structure [[1]](http://198.128.28.118/genome/converter.php#ref-1) with an e-value of 4.56647E-69 and 54% identity. Out of the active sites mentioned in the article, 17 out of 20 of them matched. The 3 mismatches were V86I, R87Q and V105I. The V86I and V105I are not major mismatches, because V and I are similar in structure. R87 and Q87 are also somewhat similar in structure, both of them having nitrogens on their side groups. Site 87 was not mentioned to have specific catalytic function either, and a reverse BLAST search gave this gene as the top hit, so this gene is predicted to code for the HisF subunit.

**Imidazoleglycerol-phosphate dehydratase**

Gene OID: 2505167253 EC: 4.2.1.19

This gene is predicted to code for imidazoleglycerol-phosphate dehydratase. A NCBI BLAST search gave only imidazoleglycerol-phosphate dehydratases as hits [[10]](http://198.128.28.118/genome/converter.php#ref-10) with e-values from 4E-115 to 3E-75 and 57% to 75% identity. A PDB BLAST search gave an article on imidazoleglycerol-phosphate dehydratase crystal structure as a top hit [[7]](http://198.128.28.118/genome/converter.php#ref-7) with an e-value of 3.32388E-72 and 65% identity. 19 out of the 20 mentioned active sites in the article matched, with the sole mismatch being H145E. H145 was not mentioned in the article as having specific catalytic function. Additionally, a reverse BLAST search gave this gene as the top hit, so this gene is predicted to code for the enzyme.

**Histidinol-phosphate aminotransferase**

Two genes are predicted by IMG/ER to code for this protein. After manual annotation, they are both predicted to code for histidinol-phosphate aminotransferase.

Gene OID: 2505165616 EC: 2.6.1.9

This gene is predicted to code for histidinol-phosphate aminotransferase. A NCBI BLAST search gave only histidinol-phosphate aminotransferases as hits {{NCBI: Q8YV89}} with e-values ranging from 4E-157 to 4E-50 and 33% to 63% identity. A PDB BLAST search gave as a top hit an article on crystal structure of histidinol-phosphate aminotransferase [[11]](http://198.128.28.118/genome/converter.php#ref-11) with an e-value of 1.08378E-39 and 30% identity. 12 out of 15 of the mentioned active sites matched up, with the 3 mismatches being N99D, H126Y and N172S. These residues were not mentioned in the article as having specific catalytic function. Furthermore, the mismatched residues are similar. N99 and D99 are very similar, just differing by a NH2 or OH group. H126 and Y126 are similar in that they both have aromatic rings. N172 and S172 are similar in that they are both polar and neutral amino acids. From this strong body of evidence, it is predicted that the gene codes for histidinol-phosphate aminotransferase.

Gene OID: 2505167361 EC: 2.6.1.9

This gene is also predicted to code for histidinol-phosphate aminotransferase. A NCBI BLAST search gave only histidinol-phosphate aminotransferases as hits {{NCBI: Q8YMG7}} with a range of e-values from 2E-179 to 3E-36 and 27% to 63% identity. A PDB BLAST search gave as a top hit the same article on histidinol-phosphate aminotransferase [[11]](http://198.128.28.118/genome/converter.php#ref-11) with an e-value of 2.23003E-39 and 34% identity. Out of the 15 active sites mentioned in the article, 11 of them matched up, with the mismatches being N99D, Y123F, H126Y and N172S. Three of these mismatches were the same as for the previous gene and were already justified. For the site 123 mismatch, the article did not mention Y123 as being important for catalytic function, and Y123 and F123 are very similar, only differing by a hydroxyl group on the aromatic benzene ring. Additionally, a reverse BLAST search gives this gene as the top hit, so this gene is predicted to code for histidinol-phosphate aminotransferase.

**Histidinol-Phosphate Phosphatase**

Gene OID: 2505167253 EC: 3.1.3.15

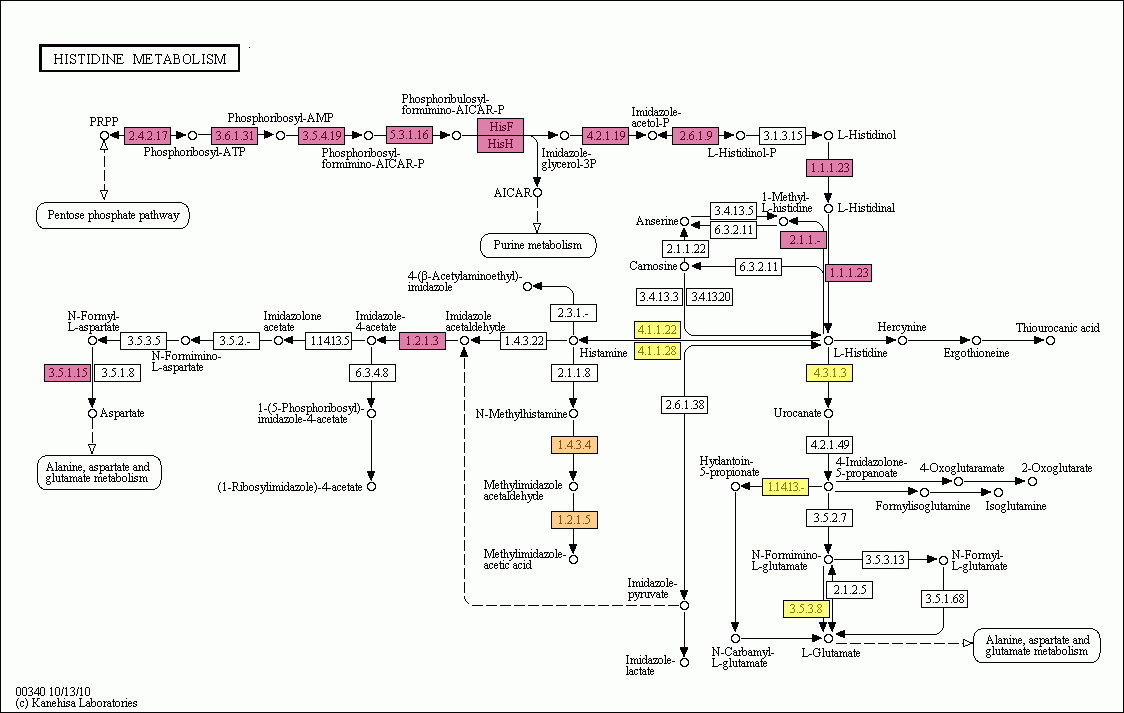
There was no gene predicted by IMG/ER to code for this protein. However, it is written in the literature [[6]](http://198.128.28.118/genome/converter.php#ref-6) that in many organisms, there is a bifunctional protein that has functionality for EC 4.2.1.19 and EC 3.1.3.15. A search was done on UniProt for the FASTA sequences of such proteins, and these reviewed hits were BLASTed against the *Microcoleus vaginatus* genome. This BLAST search gave gene 2505167253 as the top hit. This gene was previously annotated as corresponding to imidazoleglycerol-phosphate dehydratase. However, this BLAST hit suggests that the gene corresponds to a bifunctional enzyme like the ones mentioned in the literature. It is predicted that this gene is therefore bifunctional and corresponds to EC 3.1.3.15, although this conclusion needs to be double-checked in the lab.

**Histidinol dehydrogenase**

Gene OID: 2505171166 EC: 1.1.1.23

This gene is predicted to code for histidinol dehydrogenase. A NCBI BLAST search gave only other histidinol dehydrogenases as hits {{NCBI: Q8YWL4}} with e-values ranging from 0 to 6E-109 and 44% to 78% identity. A PDB BLAST search gave as a top hit an article on histidinol dehydrogenase structure [[2]](http://198.128.28.118/genome/converter.php#ref-2) with an e-value of 6.19003E-78 and 42% identity. Out of the 23 active sites mentioned in the article, 21 of them matched. The 2 mismatches were F139P and F213Y. The F139 residue was not mentioned in the article to have specific catalytic purpose. F139 and P139 are at least similar in that they are both nonpolar and neutral amino acids. The F213 residue however was mentioned as important for its aromatic ring, which Y213 also has. Also, a reverse BLAST search gave this gene as the first hit, so it is predicted that the gene codes for histidinol dehydrogenase.

**Comparison with Other Genomes**





S Fig. 2: Kegg Pathway Map of Histidine Biosynthesis for comparison with other genomes

Referring to S Figure 2, histidine biosynthesis corresponds to the topmost part of the pathway map. Looking at this pathway map, it can be seen that formost other cyanobacteria, the genes necessary for histidine biosynthesis were found, except for the gene coding for EC 3.1.3.15. This suggests that cyanobacteria in general have the bifunctional EC 3.1.3.15 and EC 4.2.1.19 protein, since they don’t have the monofunctional EC 3.1.3.15 enzyme.

For histidine biosynthesis, *Microcoleus vaginatus* is not very different from its other related cyanobacteria. There are no new genes in other cyanobacteria not found in *Microcoleus vaginatus*, and there are no new genes in *Microcoleus vaginatus* not found in other cyanobacteria.

**Functional Studies**

Almost all of the enzymes associated with the histidine biosynthesis pathway were found to have genes coding for them. The one exception was histidinol-phosphate phosphatase (EC 3.1.3.15). This enzyme was not predicted to have a gene coding for it, and it was hypothesized that the gene coding for EC 4.2.1.19, gene 2505167253, actually coded for a bifunctional protein that had function for ECs 4.2.1.19 and 3.1.3.15. An experiment needs to be done if this is actually the case; that is, if the gene actually codes for a bifunctional protein. It is known from experience that *Microcoleus vaginatus* can synthesize its own histidine, because it can grow without being given any amino acid supplements. Therefore, either gene 2505167253 codes for a bifunctional protein or there is an alternate pathway or protein that allows *Microcoleus vaginatus* to synthesize histidine.

**Discussion**

*Microcoleus vaginatus* is predicted to be able to synthesize its own histidine, and this is known to be true by experiment. The bug is missing the enzyme histidinol-phosphate phosphatase (EC 3.1.3.15) in the pathway, but it is predicted that gene 2505167253, predicted to code for EC 4.2.1.19, actually codes for a bifunctional protein with both EC 4.2.1.19 and EC 3.1.3.15 function. In this way, *Microcoleus vaginatus* can synthesize histidine.

The bifunctional protein is the most likely solution to the problem of not having a monofunctional histidinol-phosphate phosphatase enzyme. However, there is always a possibility that *Microcoleus vaginatus* uses an alternate enzyme or pathway to synthesize histidine. This should be looked into, especially if it is found that the protein is not actually bifunctional and no protein has EC 3.1.3.15 function.

Finally, it is interesting that *Microcoleus vaginatus* has two genes, 2505165616 and 2505167361, predicted to code for histidinol-phosphate aminotransferease. It should be seen if *Microcoleus vaginatus* can survive with only one of these proteins expressed, or if it needs both of them expressed in order to stay alive. It should be also be seen if there is some evolutionary advantage to having two genes coding for proteins with the same function, and whether or not these genes are expressed or repressed under certain external conditions.

**References**

[1] Banfield MJ, Lott JS, Arcus VL, McCarthy AA, Baker EN. "Structure of HisF, a histidine

biosynthetic protein from Pyrobaculum aerophilum." Acta Crystallogr D Biol Crystallogr.

2001 Nov; 57(Pt 11):1518-25.

[2] Barbosa JA, Sivaraman J, Li Y, Larocque R, Matte A, Schrag JD, Cygler M. "Mechanism of

action and NAD+-binding mode revealed by the crystal structure of L-histidinol

dehydrogenase." Proc Natl Acad Sci U S A. Feb 19, 2002; 99(4):1859-64.

[3] Champagne KS, Sissler M, Larrabee Y, Doublié S, Francklyn CS. "Activation of the

hetero-octameric ATP phosphoribosyl transferase through subunit interface

rearrangement by a tRNA synthetase paralog." J Biol Chem. Oct 7, 2005;

280(40):34096-104.

[4] Dobritzsch D, König S, Schneider G, Lu G. "High resolution crystal structure of pyruvate

decarboxylase from Zymomonas mobilis. Implications for substrate activation in pyruvate

decarboxylases." J Biol Chem. Aug 7, 1998; 273(32):20196-204.

[5] Due AV, Kuper J, Geerlof A, von Kries JP, Wilmanns M. "Bisubstrate specificity in

histidine/tryptophan biosynthesis isomerase from Mycobacterium tuberculosis by active

site metamorphosis." Proc Natl Acad Sci U S A. Mar 1, 2011; 108(9):3554-9.

[6] Fani R, Mori E, Tamburini E, Lazcano A. "Evolution of the structure and chromosomal

distribution of histidine biosynthetic genes." Orig Life Evol Biosph. 1998 Oct;

28(4-6):555-70.

[7] Glynn SE, Baker PJ, Sedelnikova SE, Davies CL, Eadsforth TC, Levy CW, Rodgers HF,

Blackburn GM, Hawkes TR, Viner R, Rice DW. "Structure and mechanism of

imidazoleglycerol-phosphate dehydratase." Structure. 2005 Dec; 13(12):1809-17.

[8] Javid-Majd F, Yang D, Ioerger TR, Sacchettini JC. "The 1.25 A resolution structure of

phosphoribosyl-ATP pyrophosphohydrolase from Mycobacterium tuberculosis." Acta

Crystallogr D Biol Crystallogr. 2008 Jun; 64(Pt 6):627-35.

[9] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[10] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N,

Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda

M, Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[11] Marienhagen J, Sandalova T, Sahm H, Eggeling L, Schneider G. "Insights into the structural

basis of substrate recognition by histidinol-phosphate aminotransferase from

Corynebacterium glutamicum." Acta Crystallogr D Biol Crystallogr. 2008 Jun; 64(Pt

6):675-85.

[12] Omi R, Mizuguchi H, Goto M, Miyahara I, Hayashi H, Kagamiyama H, Hirotsu K. "Structure

of imidazole glycerol phosphate synthase from Thermus thermophilus HB8: open-closed

conformational change and ammonia tunneling." J Biochem. 2002 Nov; 132(5):759-65.

[13] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

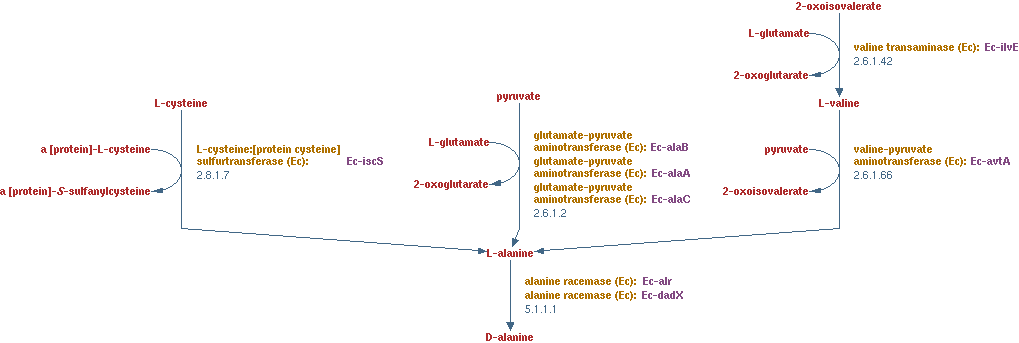
Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

## Alanine

L-alanine is an essential component of protein and peptidoglycan. The latter also contains about three molecules of D-alanine for every L-alanine. Only about 10 percent of the total alanine synthesized flows into peptidoglycan.

At least three pathways (alanine biosynthesis I, alanine biosynthesis II, and alanine biosynthesis III) contribute to the synthesis of alanine. Alanine biosynthesis I is established only by existence of the relevant enzymes. Its contribution to alanine synthesis remains speculative because alanine auxotrophs have not yet been isolated. Because alanine but not valine represses AvtA, its primary purpose is probably synthesis of L-alanine. Existence of the alanine biosyntheis II pathways rests on the evidence that lutamate-pyruvate aminotransferase acitivity is found in crude cell extracts; the enzymes has not been purified nor have mutant alleles of its designated encoding gene (alaB) been isolated. The conversion can also be mediated as a side reaction of alanine racemase. The alanine biosyntheis III pathway, mediated by cysteine desulfurase activity is required to donate sulfane sulfur for the synthesis of Fe-S clusters, thiamine, thionucleosides in tRNAs, biotin, lipoic acid, and molybdopterin probably contributes only a minor amount of the cell's alanine requirement, as judged by the cell's total requirement for sulfane sulfur.

**Automated Annotation**



S Fig. 1: MetaCyc Pathway map of alanine biosynthesis. Alanine biosynthesis I is on the right and alanine biosynthesis III is on the left, with alanine biosynthesis II in the middle.

**Valine-pyruvate aminotransferase**

Gene OID: 2505165888 EC: 2.6.1.66

This gene is predicted to code for valine-pyruvate aminotransferase. A NCBI BLAST gave a valine-pyruvate aminotransferase hit as the very top hit [[3]](http://198.128.28.118/genome/converter.php#ref-3) with an e-value of 4E-95 and 36% identity. The other hits were to other proteins, but the best hit had a high e-value of 2E-19. A PDB BLAST search gave as the top hit an unpublished article on valine-pyruvate aminotransferase (PDB: 3G7Q) with an e-value of 3.27075E-76 and 36% identity. A search was done on PDB, and this article was the only article on valine-pyruvate aminotransferase. From the strength of this PDB BLAST hit (the next best hit had an e-value of 7.15641E-23) and the strength of the NCBI BLAST hit, it is predicted that this gene codes for valine-pyruvate aminotransferase.

**Alanine racemase**

Gene OID: 2505169659 EC: 5.1.1.1

This gene is predicted to alanine racemase. A NCBI BLAST search gave only alanine racemases as hits [[4]](http://198.128.28.118/genome/converter.php#ref-4), with the e-values ranging from 0 to 7E-55 and 35% to 71% identity. A PDB BLAST search gave an article on alanine racemase as a top hit [[6]](http://198.128.28.118/genome/converter.php#ref-6) with an e-value of 4.55992E-64 and 41% identity. Out of the active sites mentioned in the article, 9 out of 9 of them matched. Additionally, a reverse BLAST search gave this gene as the top hit. Therefore, this gene is strongly predicted to code for alanine racemase.

**Cysteine sulfurtransferase**

There were three gene predicted by IMG/ER to code for this protein.These three genes were all confirmed and predicted to code for cysteine sulfurtransferase, but with differeing degrees of certainty. Gene 2505169666 is predicted to code for the protein, gene 2505168898 is possibly predicted to code for the protein, and gene 2505166725 is strongly predicted to code for the protein.

Gene OID: 2505169666 EC: 2.8.1.7

This gene is predicted to code for cysteine sulfurtransferase. A NCBI BLAST search gave only cysteine sulfurtransferases as hits [[5]](http://198.128.28.118/genome/converter.php#ref-5) with e-values from 5E-149 to 1E-138 and around 53% to 56% identity. A PDB BLAST search gave an article on cysteine sulfurtransferase as a top hit [[2]](http://198.128.28.118/genome/converter.php#ref-2) with an e-value of 2.30391E-113 and 55% identity. Of the active sites mentioned in the article, 16 out of 20 of them matched. The mismatches were S10H, T76S, K105N and Y337H. The article does not mentioned anything special about the S10 site. T76 is mentioned as important for its side-chain hydrogen bonding. S76 can hydrogen bond with its side chain as well; in fact, T76 and S76 are very similar, only differing by an extra carbon backbone on the hydroxyl side chain. K105 was mentioned as being important for helping to create a polar binding pocket for the substrate; N105 is also polar. Y337 was also mentioned as being important for creating a polar binding pocket, and H337 is also polar. Additionally, a reverse BLAST search gave this gene as the top hit. From these data, this gene is strongly predicted to code for cysteine sulfurtransferase.

Gene OID: 2505168898 EC: 2.8.1.7

This gene is also predicted to code for cysteine sulfurtransferase. A NCBI BLAST search gave only cysteine sulfurtransferases as hits [[1]](http://198.128.28.118/genome/converter.php#ref-1) with e-values from 1E-108 to 5E-87 and around 39% to 45% identity. A PDB BLAST search gave the same article on cysteine sulfurtransferase [[2]](http://198.128.28.118/genome/converter.php#ref-2), with an e-value of 4.3368E-75 and 41% identity. Out of the 20 active sites mentioned in the article, 12 of them matched. The 8 mismatches were K105S, D109E, N154Q, N155S, H231G, H235Q, L244Q and Y337P. K105 and N155 were mentioned in the article as being important for forming a polar binding environment; S105 and S155 can also accomplish this since they are also polar. D109 is not mentioned at all in the article and as such is not an important residue, and neither were N154, H231 or H235. Y337 was mentioned as important for its polarity, but P337 is not polar. To double-check this mismatch and the residues that were not mentioned in the article, a multiple sequence alignment was done on T-Coffee using cyanobacterial FASTA sequences from UniProt for comparison.

This alignment showed that at site 109, a few other cyanobacteria had the D109 residue, but most had the V109 residue. None had the E109 residue. At site 154, one other cyanobacteria had the Q154 residue. The most common residues were N154 and S154 though. At site 231, one other cyanobacteria had the H231 residue, while none had the G231 residue. The most common residues were F231 and L231. For site 235, no other cyaonbacteria had the H235 residue, but a good amount had the Q235 residue. The majority of the residues were E235. At site 337, one other cyanobacteria had the P337 residue, none had the Y337 residue, and most had the H337 residue.

The multiple sequence alignment is somewhat inconclusive, but nonetheless the gene is still predicted to code for cysteine sulfurtransferase, although this should be double checked with an experiment.

Gene OID: 2505166725 EC: 2.8.1.7

This gene is predicted to code for cysteine sulfurtransferase as well. A NCBI BLAST search gave other cysteine sulfurtransferases as top hits [[7]](http://198.128.28.118/genome/converter.php#ref-7) with e-values ranging from 0 to 5E-24 and 28% to 68% identity. A PDB BLAST search gave an article on cysteine sulfurtransferase crystal structure as a top hit [[7]](http://198.128.28.118/genome/converter.php#ref-7) with an e-value of 1.05227E-167 and 68% identity. 14 active sites were mentioned in the article, and all 14 of them matched. Also, a reverse BLAST search gave this gene as the top hit. Therefore, this gene is strongly predicted to code for cysteine sulfurtransferase.

**Glutamate-pyruvate aminotransferase**

EC: 2.6.1.2

No gene was predicted by IMG/ER to code for this protein. A search was done in IMG/ER to see if other cyanobacteria had genes for this protein. No other cyanobacteria were predicted to have such genes. A search was done on UniProt as well, with the same result. Therefore, it is predicted that *Microcoleus vaginatus* doesn’t express glutamate-pyruvate aminotransferase.

**Functional Studies**

Functional studies should be done on gene 2505168898. This gene is predicted to code for cysteine sulfurtransferase. However, only 12 out of 20 active sites matched, so this puts some doubt on whether or not the gene actually codes for cysteine sulfurtransferase. An experiment should be done where all other genes that code for cysteine sulfurtransferase are repressed, and it should be seen if cysteine sulfurtransferase is still expressed when this gene is present.

Functional studies should also be done on genes 2505166725 and 2505169666 along with gene 2505168898. These genes are all predicted to code for cysteine sulfurtransferase. It is interesting that 3 genes would all code for proteins with the same function. Experiments should be done to see if this confers some sort of evolutionary advantage for the organism, or if different genes are expressed under different external conditions. Finally, it should be seen if there is a benefit to such redundancy; perhaps this gene is often mutated or transcribed improperly, so it would make sense to have triple redundancy for as important a function as alanine biosynthesis.

**Discussion**

*Microcoleus vaginatus* is predicted to be able to synthesize both enantiomers of alanine. It is predicted to be able to synthesize L-alanine via two pathways: the cysteine pathway and the valine pathway. It is not predicted to be able to synthesize L-alanine from the pyruvate pathway. It is predicted to be able to synthesize D-alanine from L-alanine with the help of alanine racemase.

This conclusion is consistent about what is known about *Microcoleus vaginatus* from experiments. *Microcoleus vaginatus* is predicted to be able to synthesize its own amino acids, because it can grow without being supplied amino acids.

**References**

[1] Borup B, Ferry JG. "Cysteine biosynthesis in the Archaea: Methanosarcina thermophila

utilizes O-acetylserine sulfhydrylase." FEMS Microbiol Lett. Aug 15, 2000;

189(2):205-10.

[2] Cupp-Vickery JR, Urbina H, Vickery LE. "Crystal structure of IscS, a cysteine desulfurase

from Escherichia coli." J Mol Biol. Jul 25, 2003; 330(5):1049-59.

[3] Hayashi K, Morooka N, Yamamoto Y, Fujita K, Isono K, Choi S, Ohtsubo E, Baba T, Wanner

BL, Mori H, Horiuchi T. "Highly accurate genome sequences of Escherichia coli K-12

strains MG1655 and W3110." Mol Syst Biol. 2006; 2():2006.0007.

[4] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[5] Makino K, Oshima K, Kurokawa K, Yokoyama K, Uda T, Tagomori K, Iijima Y, Najima M,

Nakano M, Yamashita A, Kubota Y, Kimura S, Yasunaga T, Honda T, Shinagawa H,

Hattori M, Iida T. "Genome sequence of Vibrio parahaemolyticus: a pathogenic

mechanism distinct from that of V cholerae." Lancet. Mar 1, 2003; 361(9359):743-9.

[6] Stamper GF, Morollo AA, Ringe D. "Reaction of alanine racemase with

1-aminoethylphosphonic acid forms a stable external aldimine." Biochemistry. Jul 21,

1998; 37(29):10438-45.

[7] Tirupati B, Vey JL, Drennan CL, Bollinger JM Jr. "Kinetic and structural characterization of

Slr0077/SufS, the essential cysteine desulfurase from Synechocystis sp. PCC 6803."

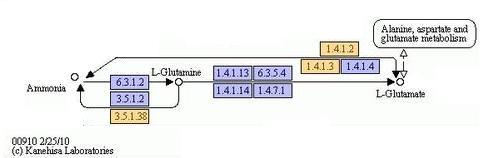
Biochemistry. Sep 28, 2004; 43(38):12210-9.

## Glutamine & Glutamate

Although atmospheric nitrogen (N2) is most likely not fixed by *M. vaginatus*, the cyanobacteria still metabolises nitrogen in the form of ammonia for its necessary functions. Ammonium may be one of those key nitrogen sources. In this annotation IMG/ER automated predictions as well as manual predictions of key enzymes, transport proteins, and regulators are studied.

Ammonium, the most reduced form of nitrogen, is a crucial step in the nitrogen assimilation process. “In cyanobacteria, after transport by speciﬁc permeases, ammonium is incorporated into carbon skeletons by the sequential action of glutamine synthetase (GS) and glutamate synthase (GOGAT). Two types of GS (GSI and GSIII) and two types of GOGAT (ferredoxin-GOGAT and NADH–GOGAT) have been characterized in cyanobacteria. The carbon skeleton substrate of the GS–GOGAT pathway is 2-oxoglutarate that is synthesized by the isocitrate dehydrogenase (IDH). In order to maintain the C–N balance and the amino acid pools homeostasis, ammonium assimilation is tightly regulated. The key regulatory point is the GS, which is controlled at transcriptional and posttranscriptional levels. The transcription factor NtcA plays a critical role regulating the expression of the GS and the IDH encoding genes. In the unicellular cyanobacterium Synechocystis sp. PCC 6803, NtcA controls also the expression of two small proteins (IF7 and IF17) that inhibit the activity of GS by direct protein–protein interaction. Cyanobacteria perceive nitrogen status by sensing the intracellular concentration of 2-oxoglutarate, a signaling metabolite that is able to modulate allosterically the function of NtcA, in vitro. In vivo, a functional dependence between NtcA and the signal transduction protein PII in controlling NtcA dependent genes has been also shown.” [[11]](http://198.128.29.183/genome/converter.php#ref-11)

**Automated Annotation**



S Fig 1. Kegg Illustration of Ammonium Assimilation & glutamate synthesis pathway of Nitrogen Metabolism. IMG/ER automated predictions are indicated by color: blue indicates a gene prediction in *M. vaginatus*, orange indicates a gene prediction in other genomes but not in *Microcoleus,* and white indicates no prediction was made. The original pathway map can be found at<http://www.genome.jp/kegg/pathway/map/map00910.html>.

**Ammonium Transport Proteins (amt1 & amt2)**

Gene OID: 2505169989 TC: 1.A.11.2.3

This is a probable Ammonium permease encoding gene inferred from homology with known Synechocystis sp. PCC 6803 amt encoding genes found in NCBI (P54147, 58% identity, E-value 0) [[10]](http://198.128.29.183/genome/converter.php#ref-10). In addition, this gene was found in the same gene neighborhood as the predicted amt2.

Gene OID: 2505169985 TC: 1.A.11.2.3

This is a probable Ammonium permease encoding gene inferred from homology with verified amt encoding genes found in NCBI (Q07429, 49% identity, 2.00E-123) [[18]](http://198.128.29.183/genome/converter.php#ref-18). This gene was also found in the same neighborhood as the predicted amt1.

**Glutaminase**

Gene OID: 2505167437 EC 3.5.1.2

This gene is predicted to code for glutaminase due to strong homology with experimentally verified glutaminase enzymes and 100% active site conservation. The top verified homolog with EC 3.5.1.2 and active site information was found in PDB (O31465, 32% identity, E-value: 6.41E-41) [[4]](http://198.128.29.183/genome/converter.php#ref-4).

**Glutamine synthetase type I**

Gene OID: 2505167133 EC: 6.3.1.2

This gene is predicted to code for L-glutamine synthetase type I due to strong homology with experimentally verified synthetase type I enzymes and 100% active site conservation. Strong homologs and active sites coding for EC 6.3.1.2 were found in PDB (P0A590, 57% identity, E-value 9.94E-165) [[9]](http://198.128.29.183/genome/converter.php#ref-9) and Swissprot (P77961, 81% identity, E-value 0) [[15]](http://198.128.29.183/genome/converter.php#ref-15).

**Glutamine synthetase type III**

Gene OID: 2505169287 EC: 6.3.1.2

This gene is predicted to code for type-3 glutamate--ammonia ligase type III due to strong homology with experimentally verified synthetase type III enzymes and 100% active site conservation. Strong homologs and active sites coding for EC 6.3.1.2 were found in PDB (Q5LGP1, 43% identity, E-value 5.91515E-165) [[16]](http://198.128.29.183/genome/converter.php#ref-16) and Swissprot (Q54WR9, 46% identity, E-value 0) [[6]](http://198.128.29.183/genome/converter.php#ref-6).

**Glutamate synthase (NADH/NADPH) small subunit**

Gene OID: 2505168845 EC: 1.4.1.13/1.4.1.14

This gene is predicted to code for the small subunit of glutamate synthase (NADH/NADPH) due to strong homology with experimentally verified synthase enzymes. The top verified homolog with EC 1.4.1.13/1.4.1.14 was found in Swissprot (O34399, 60% identity, E-value 0) [[1]](http://198.128.29.183/genome/converter.php#ref-1).

**Glutamate synthases (NADH/NADPH) large subunit**

Gene OID: 2505169423 EC: 1.4.1.13/1.4.1.14

This gene is predicted to code for the large subunit of glutamate synthase (NADH/NADPH) due to strong homology and close active site matching with experimentally verified synthase enzymes. The top verified homolog with active site information was found in PDB (Q05755, 44% identity, E-value 0) [[2]](http://198.128.29.183/genome/converter.php#ref-2). Active site mismatches included Q1107-V, G1093-A, and G857-A out of 16 active sites. Further research may be necessary if a thorough investigation of these substitutions is crucial.

**Asparagine synthase (glutamine-hydrolysing)**

Gene OID: 2505166041 & 2505168423 EC: 6.3.5.4

These are possible Asparagine synthase (glutamine-hydrolysing) encoding genes inferred from homology. A weak homolog from *Bacillus subtilis* (P54420) [[19]](http://198.128.29.183/genome/converter.php#ref-19) was found in Swissprot (2505166041: 24% identity, E-value 7.00E-42; 2505168423: 22% identity, E-value 2.00E-27). In addition, a multiple sequence alignment with T-coffee using different asapargine synthases predicted in cyanobacteria showed that only certain domains were conserved (in terms of amino acid properties).

**Glutamate synthase (ferredoxin dependent)**

Gene OID: 2505170210 EC: 1.4.7.1

This gene is predicted to code for ferredoxin dependent glutamate synthase due to strong homology and close active site matching with experimentally verified enzymes with EC 1.4.7.1. The top verified homolog with active site information was found in PDB (P55038, 65% identity, E-value 0) [[22]](http://198.128.29.183/genome/converter.php#ref-22). The only active site mismatch (out of 16 sites) was Ser1129-T, which most likely does not drastically affect the protein’s function.

**Glutamate dehydrogenase**

Gene OID: 2505169473 EC: 1.4.1.3

This is a probable GDH encoding gene inferred from homology and active site conservation. Strong homologs with EC 1.4.1.2, 1.4.1.3 and 1.4.1.4 were found in NCBI and PDB databases, which suggests that the protein can most likely handle both NAD+ and NADP+ as cofactors. Active sites matched 100% in two different homologs in PDB with EC 1.4.1.2 (O74024, 46% identity, E-value 1.28E-102) [[13]](http://198.128.29.183/genome/converter.php#ref-13)and 1.4.1.3 (P96110, E=3.54703E-91, 43% identity) [[8]](http://198.128.29.183/genome/converter.php#ref-8).

**global nitrogen regulator NtcA**

Gene OID: 2505167256

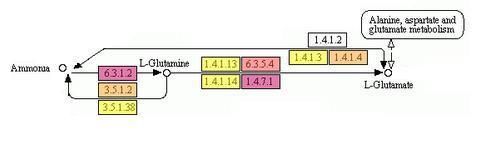
This is a probable global nitrogen regulator NtcA inferred from homology with experimentally verified homologs found in NCBI (P0A4U6, 87% identity, E-value 1.00E-141) [[17]](http://198.128.29.183/genome/converter.php#ref-17). In addition to this homology, 2-oxoglutarate binding site conservation was observed with another homology in PDB (P0A4U6) [[20]](http://198.128.29.183/genome/converter.php#ref-20).

**Nitrogen regulatory protein PII**

Gene OID: 2505170043

This is a probable PII encoding gene inferred from homology and binding site (with 2-oxoglutarate) conservation. Strong homologs were found in PDB (O30794, 92% identity, E-value 2.00E-69) [[7]](http://198.128.29.183/genome/converter.php#ref-7) as well as known binding sites (P0A3F4) [[5]](http://198.128.29.183/genome/converter.php#ref-5), which were conserved in this predicted protein.

**Comparison with other genomes**





S Fig 2. IMG/ER illustration of enzyme conservation among 68 different cyanobacterial genomes in Kegg pathway.

As a result of comparisons with other genomes, enzymes encoding Glutamin-(asparagin-)ase EC 3.5.1.38 most likely do not exist in *M. vaginatus* due to poor homology (via BLAST) between known Glutamin-(asparagin-)ase encoding genes and genes in *Microcoleus*. For a discussion of enzyme’s expressing Glutamate dehydrogenase (EC 1.4.1.2, 1.4.1.3, 1.4.1.4) look to the annotation of gene OID: 2505169473.

In addition to genes not predicted to IMG, literature suggested regulatory proteins such as PipX, a PII interaction protein, [[12]](http://198.128.29.183/genome/converter.php#ref-12)and glutamine synthase inactivating factors IF7 & IF17 [[12]](http://198.128.29.183/genome/converter.php#ref-12) found in *Synechocystis sp.* PCC 6803 were annotated as possible genes in *Microcoleus*.

**Nitrogen regulatory protein, PipX (PII interaction protein X)**

Gene OID: 2505167809

This is a probable PipX protein encoding gene inferred from homology and binding site conservation. Strong homologs were found in PDB (Q8YZH5, 62% identity, E-value 8.01E-29) [[21]](http://198.128.29.183/genome/converter.php#ref-21).

**Glutamine synthase inactivating factors IF7 & IF17**

Gene OID: 2505168417

This is a possible IF7 encoding gene inferred from homology with a known IF7 protein ( genbank: BAA17150, 72% identity, E-value 4.00E-21) [[12]](http://198.128.29.183/genome/converter.php#ref-12) around the same length (65-68 aa). Because IF7 is only a suspected posttranscriptional inhibitor of GSI nitrogen metabolism and not well studied further research is necessary to verify if this protein is in fact a GSI inhibitor.

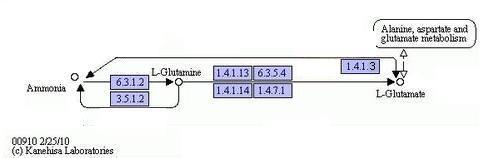
Gene OID: 2505168586

This is a possible IF17 encoding gene inferred from weak homology with a verified IF17 protein (genbank: BAA16994, 27% identity, E-value 1.00E-05) [[12]](http://198.128.29.183/genome/converter.php#ref-12). Because IF17 is only a suspected posttranscriptional inhibitor of GSI nitrogen metabolism and is not well studied, this requires further research to verify.

**Functional Studies**

Functional studies on genes suspected to encode the IF7 & IF17 inhibitors and Asparagine synthase (glutamine-hydrolysing) may be necessary in the future because their function is relatively uncertain due to weak homology. In a past study [[14]](http://198.128.29.183/genome/converter.php#ref-14) the glutamine synthetase of the cyanobacterium *Synechocystis sp.* PCC 6803 could be inactivated in vivo by ammonium addition, and the presence of the inhibitors were found by construction of a Synechocystis strain harboring a histidine-tagged modiﬁed version of GSI, allowing the puriﬁcation of the inactive enzyme and identification of inhibiting polypeptides. A similar study could be done on *Microcoleus*. Asparagine synthase function could be tested by constructing mutant strains of Microcoleus lacking the possible genes and observing what affect that has on growth in the presence of asparagine, which was done in a study on *Bacillus subtilis* [[19]](http://198.128.29.183/genome/converter.php#ref-19).

**Final Annotation**



S Fig. 3 Final annotated pathway map for *Microcoleus Vaginatus* PCC 9802 derived from IMG/ER predictions and other genome comparisons. Displays predicted enzymes.

Genes encoding ammonium transport proteins, enzymes involved in glutamiine and glutamate synthesis like GSI, GSII, and GDH, and nitrogen regulatory proteins like NtcA and PII/PipX (which are referenced throughout nitrogen metabolic pathways) are predicted and annotated within *Microcoleus vaginatus* PCC 9802. *Microcoleus* most likely assimilates ammonia as a nitrogen source.

**Discussion**

In light of the fact that not much research has been found about *Microcoleus’s* nitrogen capabilities, this gene annotation may provide insight into how it the cyanobacteria survives in the desert. For example, GDH is not strictly conserved in cyanobacteria, but because it does not require ATP it may be a more efficient pathway for glutamate synthesis when energy is scarce. In addition, GDH has been proven to be very thermostable in some bacteria [[3]](http://198.128.29.183/genome/converter.php#ref-3), a property that would be useful in the desert. GSIII, which is also not strictly conserved, has also been suggested to be more efficient than GSI in nitrogen deficient situations [[11]](http://198.128.29.183/genome/converter.php#ref-11), which could be another adaptive property in the desert, and perhaps during dormant periods. In the end, this annotation also serves as one of the key building blocks to constructing a clear picture of how *Microcoleus* metabolizes nitrogen.

**References**

[1] Belitsky BR, Wray LV Jr, Fisher SH, Bohannon DE, Sonenshein AL. "Role of TnrA in nitrogen source-dependent repression of Bacillus subtilis glutamate synthase gene expression." J Bacteriol. 2000 Nov; 182(21):5939-47.

[2] Binda C, Bossi RT, Wakatsuki S, Arzt S, Coda A, Curti B, Vanoni MA, Mattevi A. "Cross-talk and ammonia channeling between active centers in the unexpected domain arrangement of glutamate synthase." Structure. Dec 15, 2000; 8(12):1299-308.

[3] Britton KL, Yip KS, Sedelnikova SE, Stillman TJ, Adams MW, Ma K, Maeder DL, Robb FT, Tolliday N, Vetriani C, Rice DW, Baker PJ. "Structure determination of the glutamate dehydrogenase from the hyperthermophile Thermococcus litoralis and its comparison with that from Pyrococcus furiosus." J Mol Biol. Nov 12, 1999; 293(5):1121-32.

[4] Brown G, Singer A, Proudfoot M, Skarina T, Kim Y, Chang C, Dementieva I, Kuznetsova E, Gonzalez CF, Joachimiak A, Savchenko A, Yakunin AF. "Functional and structural characterization of four glutaminases from Escherichia coli and Bacillus subtilis." Biochemistry. May 27, 2008; 47(21):5724-35.

[5] Fokina O, Chellamuthu VR, Forchhammer K, Zeth K. "Mechanism of 2-oxoglutarate signaling by the Synechococcus elongatus PII signal transduction protein." Proc Natl Acad Sci U S A. Nov 16, 2010; 107(46):19760-5.

[6] Gotthardt D, Blancheteau V, Bosserhoff A, Ruppert T, Delorenzi M, Soldati T. "Proteomics fingerprinting of phagosome maturation and evidence for the role of a Galpha during uptake." Mol Cell Proteomics. 2006 Dec; 5(12):2228-43.

[7] Hanson TE, Forchhammer K, Tandeau de Marsac N, Meeks JC. "Characterization of the glnB gene product of Nostoc punctiforme strain ATCC 29133: glnB or the PII protein may be essential." Microbiology. 1998 Jun; 144 ( Pt 6)():1537-47.

[8] Knapp S, de Vos WM, Rice D, Ladenstein R. "Crystal structure of glutamate dehydrogenase from the hyperthermophilic eubacterium Thermotoga maritima at 3.0 A resolution." J Mol Biol. Apr 11, 1997; 267(4):916-32.

[9] Krajewski WW, Jones TA, Mowbray SL. "Structure of Mycobacterium tuberculosis glutamine synthetase in complex with a transition-state mimic provides functional insights." Proc Natl Acad Sci U S A. Jul 26, 2005; 102(30):10499-504.

[10] Montesinos ML, Muro-Pastor AM, Herrero A, Flores E. "Ammonium/methylammonium permeases of a Cyanobacterium. Identification and analysis of three nitrogen-regulated amt genes in synechocystis sp. PCC 6803." J Biol Chem. Nov 20, 1998; 273(47):31463-70.

[11] Muro-Pastor MI, Reyes JC, Florencio FJ. "Ammonium assimilation in cyanobacteria." Photosynth Res. 2005; 83(2):135-50.

[12] Mérida A, Candau P, Florencio FJ. "In vitro reactivation of in vivo ammonium-inactivated glutamine synthetase from Synechocystis sp. PCC 6803." Biochem Biophys Res Commun. Dec 16, 1991; 181(2):780-6.

[13] Nakasako M, Fujisawa T, Adachi S, Kudo T, Higuchi S. "Large-scale domain movements and hydration structure changes in the active-site cleft of unligated glutamate dehydrogenase from Thermococcus profundus studied by cryogenic X-ray crystal structure analysis and small-angle X-ray scattering." Biochemistry. Mar 13, 2001; 40(10):3069-79.

[14] Reyes JC, Florencio FJ. "A novel mechanism of glutamine synthetase inactivation by ammonium in the cyanobacterium Synechocystis sp. PCC 6803. Involvement of an inactivating protein." FEBS Lett. Jun 19, 1995; 367(1):45-8.

[15] Reyes JC, Florencio FJ. "Electron transport controls transcription of the glutamine synthetase gene (glnA) from the cyanobacterium Synechocystis sp. PCC 6803." Plant Mol Biol. 1995 Feb; 27(4):789-99.

[16] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes." Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[17] Wei TF, Ramasubramanian TS, Pu F, Golden JW. "Anabaena sp. strain PCC 7120 bifA gene encoding a sequence-specific DNA-binding protein cloned by in vivo transcriptional interference selection." J Bacteriol. 1993 Jul; 175(13):4025-35.

[18] Wray LV Jr, Atkinson MR, Fisher SH. "The nitrogen-regulated Bacillus subtilis nrgAB operon encodes a membrane protein and a protein highly similar to the Escherichia coli glnB-encoded PII protein." J Bacteriol. 1994 Jan; 176(1):108-14.

[19] Yoshida K, Fujita Y, Ehrlich SD. "Three asparagine synthetase genes of Bacillus subtilis." J Bacteriol. 1999 Oct; 181(19):6081-91.

[20] Zhao MX, Jiang YL, He YX, Chen YF, Teng YB, Chen Y, Zhang CC, Zhou CZ. "Structural basis for the allosteric control of the global transcription factor NtcA by the nitrogen starvation signal 2-oxoglutarate." Proc Natl Acad Sci U S A. Jul 13, 2010; 107(28):12487-92.

[21] Zhao MX, Jiang YL, Xu BY, Chen Y, Zhang CC, Zhou CZ. "Crystal structure of the cyanobacterial signal transduction protein PII in complex with PipX." J Mol Biol. Sep 24, 2010; 402(3):552-9.

[22] van den Heuvel RH, Ferrari D, Bossi RT, Ravasio S, Curti B, Vanoni MA, Florencio FJ, Mattevi A. "Structural studies on the synchronization of catalytic centers in glutamate synthase." J Biol Chem. Jul 5, 2002; 277(27):24579-83.

## Arginine

Arginine biosynthesis is notable for its complexity and variability at the genetic level, and by its connection with several other pathways, such as pyrimidine and polyamine biosynthesis, and certain degradative pathways. The initial steps of the arginine biosynthetic pathways proceed via *N*-acetylated intermediates. The presumed reason for this is that the acetylation prevents the spontaneous cyclization of glutamate derivatives, which leads to proline biosynthesis, thus keeping the two pathways separate.

At least three alternative arginine biosynthetic pathways have evolved in bacteria:

In the first pathway, which is not very common, the key intermediate *N*-acetyl-L-glutamate (NAG) is formed by the enzyme N-acetylglutamate synthase. A second key intermediate downstream, *N*-acetyl-L-ornithine, is hydrolyzed by the enzyme acetylornithine deacetylase, releasing acetate, and forming the arginine precursor L-ornithine.

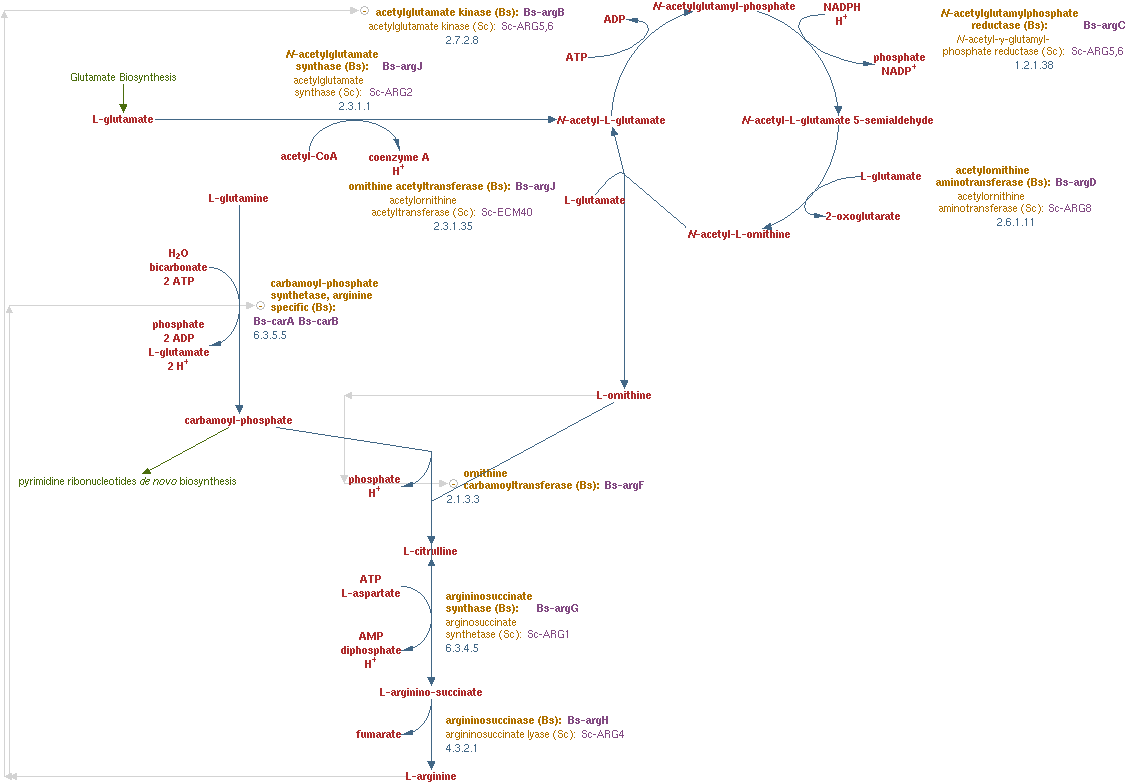
In the second pathway, which is found in most of the prokaryotic and eukaryotic microbes investigated up to now, the two reactions mentioned above are linked: the acetyl group which is removed from *N*-acetyl-L-ornithine is recycled onto glutamate, regenerating *N*-acetyl-L-glutamate (NAG). This recycling is performed by the enzyme ornithine acetyltransferase / *N*-acetylglutamate synthase (OAT). This pathway is considered more evolved, since the overall reaction is energetically more favorable.

The third pathway is a variation found in several eubacteria. While in the two pathways mentioned above *N*-acetyl-L-ornithine is deacetylate to L-ornithine, which is subsequntly transcarbamylated to form L-citrulline, in this variant *N*-acetyl-L-ornithine is not deacetylated. Instead, it is transcarbamylated directly by the enzyme acetylornithine transcarbamylase, resulting in *N*-acetyl-L-citrulline. The enzyme acetylornithine deacetylase can accept *N*-acetyl-L-citrulline as a substrate, and deacetylates it into L-citrulline.

*Microcoleus vaginatus* was found to have the enzymes for the second pathway. This cyclic pathway for arginine biosynthesis has been demonstrated in many organisms, including the prokaryotes *Bacillaceae*, pseudomonads, cyanobacteria, photosynthetic bacteria and archaebacteria, as well as eukaryotic organisms such as *Saccharomyces cerevisiae, Neurospora* sp., and *Chlamydomonas* sp. Organisms that employ this pathway possess the *argJ* gene, which encodes the enzyme ornithine acetyltransferase (OAT). OAT catalyzes the transfer of the acetyl group from *N*-α-acetyl-ornithine to glutamate, generating ornithine and *N*-acetyl-glutamate (NAG). Since the later is the substrate for another enzyme in this pathway, *N*-acetyl-glutamate kinase (encoded by *argB*), which catalyzes a previous step, this reaction effectively converts the otherwise linear pathway to a cyclical pathway. While most of the NAG in these organisms is generated by OAT, some of it is still generated by *N*-acetyl-ornithine synthase (NAGS, EC 2.3.1.1, encoded by *argA*), which fulfils an anaplerotic (replenishment of pathway intermediates) role to replenish NAG that is lost due to degradation or cell division.

In some bacteria, such as *Bacillus* sp., *Nisseria gonorrhoeae*, *Thermotoga neapolitana* and in yeast (*Saccharomyces cerevisiae*), OAT can have a dual enzymic function. In these organisms, either *N*-acetylornithine or acetyl-CoA can be used as acetyl donors for formation of NAG. *argJ* genes from such organisms can complement both *argE* and *argA* mutations in *E. coli*. However, these organisms still possess separate NAGS enzymes encoded by *argA* genes. The biological role of duplicate NAGS activity in these organisms remains to be elucidated.

**Automated Annotation**



S Fig. 1: MetaCyc pathway map of arginine biosynthesis (acetyl cycle). Original image can be found at <http://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=ARGSYNBSUB-PWY&detail-level=2>

**Argininosuccinate lyase**

Gene OID: 2505168967 EC: 4.3.2.1

This gene is predicted to code for argininosuccinate lyase. A NCBI BLAST search gave only argininosuccinate lyases as very strong hits [[12]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-12) with e-values from 0 to 6E-166 and 54% to 83% identity. A PDB BLAST search gave as a top hit an article on argininosuccinate lyase crystal structure [[1]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-1) with an e-value of 5.94794E-111 and 46% identity. This article mentioned 9 active sites, and all 9 of these mentioned active sites matched. A reverse BLAST search gave this gene as the top hit. This gene is therefore strongly predicted to code for argininosuccinate lyase.

**Argininosuccinate synthase**

Gene OID: 2505168303 EC: 6.3.4.5

This gene is predicted to code for argininosuccinate synthase. A NCBI BLAST search gave only argininosuccinate synthases as very strong hits [[9]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-9) with e-values from 0 to 2E-145 and 52% to 83% identity. A PDB BLAST search gave as a top hit an article on the crystal structure of argininosuccinate synthase [[2]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-2) with an e-value of 8.26943E-114 and 51% identity. 18 active sites were mentioned in the article, and out of these 18 active sites, 16 of them matched. The two mismatches were S89A and S182A. Site 89 was not mentioned in the article as having specific catalytic function. S182 was said to be important for its ability to hydrogen bond. A182 cannot hydrogen bond. However, besides for the fact that serine is polar and alanine is not, serine and alanine are similar in structure, and it seems that there is some kind of trend because both the serine mismatches were replaced with alanines. Besides this, a reverse BLAST search gave this gene as the top hit. The data as a whole still leads one to predict that this gene codes for argininosuccinate synthase.

**Ornithine carbamoyltransferase**

Gene OID: 2505167734 EC: 2.1.3.3

This gene is predicted to code for ornithine carbamoyltransferase. A NCBI BLAST search gave only ornithine carbamoyltransferases as hits [[3]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-3), with e-values from 2E-169 to 7E-85 and 44% to 75% identity. A PDB BLAST search gave as a top hit an article on ornithine carbamoyltransferase crystal structure [[5]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-5) with an e-value of 7.42834E-77 and 48% identity. 15 active sites were mentioned in the article, and 9 of them matched. The 6 mismatchers were E64T, H73Q, Y76D, W146K, K148C and D284E. Residues E64, H73 and K148 were mentioned to form salt bridges, which are hydrogen bonding interactions coupled with electrostatic interactions. T64 is similar to E64 in that it can hydrogen bond, but it is neutral instead of negatively charged. Q73 can also hydrogen bond, but is also neutral instead of negative like H73. C148 can hydrogen bond like K148, but is also neutral, while K148 is positive. For these substrates, although the charges aren’t the same, they can at least all hydrogen bond. For Y76, this residue was mentioned as important for a cation-pi interaction, which D76 cannot contribute to because it does not have any aromatic group. Similarly, K146 is a mismatch for W146 because W146 is important for aromatic interactions. Finally, D284 was not mentioned in the article.

To clear up these ambiguities, a multiple sequence alignment was done on T-Coffee using reviewed cyanobacterial FASTA sequences from UniProt. This alignment showed that all of the mismatches were found in other cyanobacteria. At site 64, around half of the cyanobacteria had the T64, but none had the E64 residue. At site 73, most of the other cyanobacteria had the Q73 residue, while none had the H73 residue. For site 76, all the other cyanobacteria had the D76 residue and none had the Y76 residue. At site 146, there were two other cyanobacteria with the K146 residue, but none with the W146 residue. For site 148, no cyanobacteria had the K148 residue, but a few had the C148 residue. Finally, at site 284, there were no D284 residues. Rather, all the cyanobacteria had the E284 residue.

This alignment, along with a reverse BLAST search that gave this gene as the top hit, make it reasonable to predict that the gene codes for ornithine carbamoyltransferase.

**Glutamate N-acetyltransferase/amino-acid acetyltransferase**

Gene OID: 2505168419 EC: 2.3.1.1/2.3.1.35

Gene 2505168419 probably encodes a bifunctional glutamate N-acetyltransferase/amino-acid acetyltransferas due to strong homology with experimentally verified enzymes that possess EC #’s 2.3.1.1 & 2.3.1.35. A bifunctional homolog was found via BLAST in the PDB (Q3C251, 56% identity, E-value 9.00E-15) [[10]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-10). A T-coffee sequence alignment using the 2505168419 amino acid sequence and the sequences of Swissprot verified bifunctional acetyltransferases (Q8YVA8, Q8YPF9, Q7NE46, Q7VEF9, Q7V436, Q46I07, Q7U3S6, Q8DHN4) showed strong alignment across the whole sequence.

**Amino-acid N-acetyltransferase**

Gene OID: 2505169563 EC: 2.3.1.1

Gene 2505169563 possibly encodes an amino acid N-acetyltransferase fragment due to homology with experimentally verified enzymes with EC 2.3.1.1 and a significantly shorter length (156 aa) than known N-acetyltransferases (~400 aa). Homologs were found in the Swissprot database via BLAST; the alignment, as expected, was only over a short segment of the homolog’s sequence (O66143, 26% identity, E-value 0.026) [[6]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-6).

**Acetylglutamate kinase**

Gene OID: 2505170503 EC: 2.7.2.8

Gene 2505170503 probably encodes an acetylglutamate kinase due to homology with experimentally verified acetylglutamate kinases and complete active site conservation. A strong homolog with known active sites was found in PDB via BLAST (Q6V1L5, 71% identity, E-value 1.28E-121) [[4]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-4). The active sites were conserved in the Microcoleus protein.

**N-acetyl-gamma-glutamyl-phosphate reductase**

Gene OID: 2505166396 EC: 1.2.1.38

Gene 2505166396 probably encodes a N-acetyl-gamma-glutamyl-phosphate reductase due to strong homology with experimentally verified N-acetyl-gamma-glutamyl-phosphate reductases and complete active site conservation. A strong homolog that possessed known active sites was found via BLAST in PDB (Q6AV34 Y 35 1.48E-54) [[7]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-7). These active sites were conserved in the Microcoleus protein.

**Acetylornithine and succinylornithine aminotransferases**

Gene OID: 2505170763 EC: 2.6.1.11/2.6.1.17

Gene 2505170763 probably encodes an acetylornithine and succinylornithine aminotransferase due to strong homology with experimentally verified enzymes with EC 2.6.1.11 & 2.6.1.17 and close active site conservation. A strong homolog with EC’s 2.6.1.11 & 2.6.1.17 and 8 known active sites was found in PDB via BLAST (P40732, 37% identity, E-value 1.20E-81) [[8]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-8). Out of these 8 active sites 2 were not conserved; mismatches were Q229G and T109A. It is unknown if these changes significantly affect functionality.

**Carbamoyl-phosphate synthase, small subunit**

Gene OID: 2505168986 EC: 6.3.5.5

Gene 2505168986 probably encodes carbamoyl-phosphate synthase due to strong homology with experimentally verified small subunits and complete active site conservation. A strong homolog verified to be a carbamoyl-phosphate synthase small subunit with known active sites was found in PDB via BLAST (P0A6F1, 50% identity, E-value 4.01E-102) {{pubmed: 10029528}}. These active sites were conserved in the Microcoleus protein.

**Carbamoyl-phosphate synthase, large subunit**

Gene OID: 2505166825 EC: 6.3.5.5

Gene 2505166825 probably encodes carbamoyl-phosphate synthase due to strong homology with experimentally verified large subunits and complete active site conservation. A strong homolog verified to be a carbamoyl-phosphate synthase large subunit with known active sites was found in PDB via BLAST (P00968, 58% identity, E-value 0) [[11]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-11). These active sites were conserved in the Microcoleus protein.

**Discussion**

From this pathway annotation, *Microcoleus vaginatus* is predicted to be able to synthesize its own arginine. All the enzymes necessary for the acetyl cycle variation of arginine biosynthesis were predicted to be found in the bug. This result is consistent with what is known from experiment, since *Microcoleus vaginatus* is known to be able to grow without any amino acid sources, which means that it synthesizes its own amino acids. This annotation was straightforward and did not have ambiguity with genes coding for enzymes, so it can be said that *Microcoleus vaginatus* produces its own arginine.

**References**

[1] Bhaumik P, Koski MK, Bergmann U, Wierenga RK. "Structure determination and refinement

at 2.44 A resolution of argininosuccinate lyase from Escherichia coli." Acta Crystallogr D

Biol Crystallogr. 2004 Nov; 60(Pt 11):1964-70.

[2] Goto M, Omi R, Miyahara I, Sugahara M, Hirotsu K. "Structures of argininosuccinate

synthetase in enzyme-ATP substrates and enzyme-AMP product forms: stereochemistry

of the catalytic reaction." J Biol Chem. Jun 20, 2003; 278(25):22964-71.

[3] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[4] Llácer JL, Contreras A, Forchhammer K, Marco-Marín C, Gil-Ortiz F, Maldonado R, Fita I,

Rubio V. "The crystal structure of the complex of PII and acetylglutamate kinase reveals

how PII controls the storage of nitrogen as arginine." Proc Natl Acad Sci U S A. Nov 6,

2007; 104(45):17644-9.

[5] Massant J, Wouters J, Glansdorff N. "Refined structure of Pyrococcus furiosus ornithine

carbamoyltransferase at 1.87 A." Acta Crystallogr D Biol Crystallogr. 2003 Dec; 59(Pt

12):2140-9.

[6] Nakabachi A, Ishikawa H. "Differential display of mRNAs related to amino acid metabolism in

the endosymbiotic system of aphids." Insect Biochem Mol Biol. 1997 Dec;

27(12):1057-62.

[7] Nonaka T, Kita A, Miura-Ohnuma J, Katoh E, Inagaki N, Yamazaki T, Miki K. "Crystal

structure of putative N-acetyl-gamma-glutamyl-phosphate reductase (AK071544) from

rice (Oryza sativa)." Proteins. Dec 1, 2005; 61(4):1137-40.

[8] Rajaram V, Ratna Prasuna P, Savithri HS, Murthy MR. "Structure of biosynthetic

N-acetylornithine aminotransferase from Salmonella typhimurium: studies on substrate

specificity and inhibitor binding." Proteins. Feb 1, 2008; 70(2):429-41.

[9] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[10] Takahara K, Akashi K, Yokota A. "Purification and characterization of glutamate

N-acetyltransferase involved in citrulline accumulation in wild watermelon." FEBS J.

2005 Oct; 272(20):5353-64.

[11] Thoden JB, Wesenberg G, Raushel FM, Holden HM. "Carbamoyl phosphate synthetase:

closure of the B-domain as a result of nucleotide binding." Biochemistry. Feb 23, 1999;

38(8):2347-57.

[12] Troshina O, Hansel A, Lindblad P. "Cloning, characterization, and functional expression in

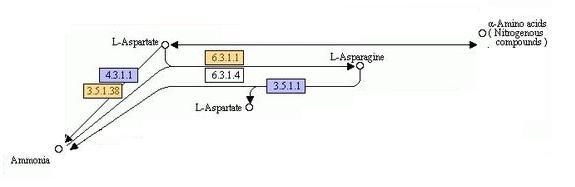
Escherichia coli of argH encoding argininosuccinate lyase in the cyanobacterium Nostoc

sp. strain PCC 73102." Curr Microbiol. 2001 Oct; 43(4):260-4.

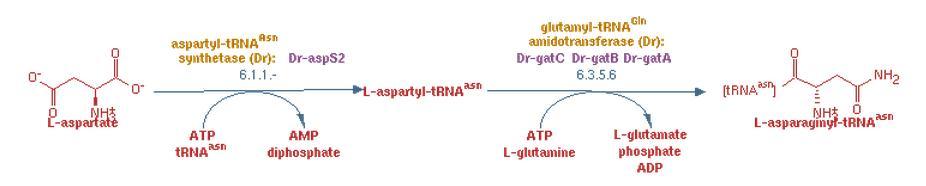
## Asparagine

In both eukaryotes and prokaryotes, L-asparagine is biosynthesized from L-aspartate by amidation using L-glutamine as an amino group donor. In some organisms, including *Escherichia coli* K-12 , a second pathway exists that utilizes ammonia instead. A third path includes tRNA dependent synthesis. In this annotation, several pathways of asparagine synthesis are explored and relevant genes annotated.

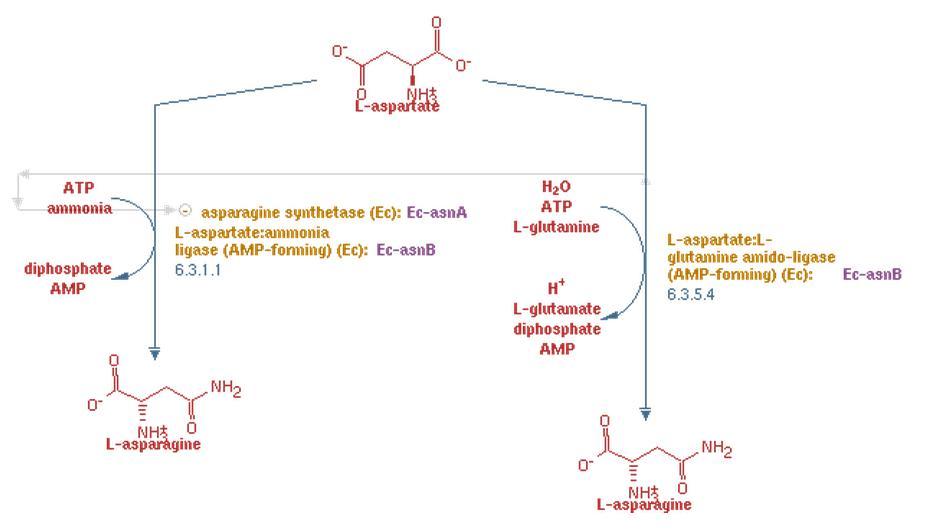
**Automated Annotation**



S Fig 1. Kegg Illustration of Asparagine synthesis in Nitrogen Metabolism Pathway. Original image can be found at <http://www.genome.jp/kegg/pathway/map/map00910.html>. IMG/ER automated predictions are indicated by color: blue indicates a gene prediction in *M. vaginatus*, orange indicates a gene prediction in other genomes but not in *Microcoleus,* and white indicates no prediction was made.



S Fig 2. Metacyc illustration of tRNA-dependent Asparagine biosynthesis.



S Fig 3. Metacyc illustration of Asparagine/Aspartate conversion pathway.

**Asparagine synthase (glutamine-hydrolysing)**

Gene OID: 2505166041 & 2505168423 EC: 6.3.5.4

These are possible Asparagine synthase (glutamine-hydrolysing) encoding genes inferred from homology. A weak homolog from *Bacillus subtilis* (P54420) [[8]](http://198.128.28.118/genome/converter.php#ref-8) was found in Swissprot (2505166041: 24% identity, E-value 7.00E-42; 2505168423: 22% identity, E-value 2.00E-27). In addition, a multiple sequence alignment with T-coffee using different asapargine synthases predicted in cyanobacteria showed that only certain domains were conserved (in terms of amino acid properties).

**Asparaginase**

Gene OID: 2505170135; 2505171163 EC: 3.5.1.1

These are probable asparginase encoding genes inferred from homology (active sites unknown). Strong experimentally verified homologs were found in via BLAST search in a Swissprot database for both gene 2505170135 (Q8YQB1, 72% identity, E-value 4.00E-163) [[3]](http://198.128.28.118/genome/converter.php#ref-3) and 2505171163 (Q8VI04, 44% identity, E-value 3.00E-61) [[2]](http://198.128.28.118/genome/converter.php#ref-2).

**Aspartase**

Gene OID 2505169957 EC 4.3.1.1, 4.2.1.2

This is a probable aspartate ammonia-lyase encoding gene inferred from homology and active site conservation. Strong experimentally verified homologs with known active sites were found in PDB (Q9LCC6, 45% identity, E-value 5.78E-114) [[6]](http://198.128.28.118/genome/converter.php#ref-6). This gene has also been annotated in previous reports as possessing EC 4.2.1.2.

**Aspartyl-tRNA synthetase**

Gene OID: 2505166347 EC: 6.1.1.12

This is a probable aspartyl-tRNA synthetase encoding gene inferred from homology and 100% active site conservation. Strong experimentally verified homologs with known active sites were found in PDB (P36419, 49% identity, E-value 1.96E-152) [[1]](http://198.128.28.118/genome/converter.php#ref-1).

**Asparaginyl-tRNA synthase, aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase**

Gene OID: 2505167906; 2505167938; 2505168995; 2505169086; 2505168184 2505168082

EC: 6.3.5.6, 6.3.5.7, 3.5.1.4

Multiple genes were predicted by IMG/ER to contain this function, among others. Three genes were predicted to code for the subunits A, B, & C due to homology and active site conservation. Gene 2505167906 possessed strong verified C subunit homologs in PDB (O67904, 46% identity, E-value 9.37E-15) [[7]](http://198.128.28.118/genome/converter.php#ref-7). Genes 2505167938 and 2505168995 found strong homology with B & C subunits, respectively, in PDB (O66766, 43% identity, E-value 4.69E-143; O66610, 54% identity, E-value 2.71E-147) [[7]](http://198.128.28.118/genome/converter.php#ref-7) that possessed active site information. Active sites on subunit A were conserved 100% and ~80% on subunit B (A49-V, V42-I were the non-conserved residues). Interestingly, the homologs were verified to have EC # 6.3.5.- and were bifunctional for aspartyl and glutamyl (tRNA dependent) amidotransferase; perhaps these enzymes in *Microcoleus* are multifunctional as well.

The other genes IMG predicted to encode aspartyl-tRNA amidotransferase (2505169086, 2505168184, and 2505168082) were assigned multiple functions: aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase and amidase (EC 6.3.5.6/7 & 3.5.1.4). Although they all beared some homology to the transferase subunit A proteins, they seemed to be only fragments and existed at the ends of gene scaffolds, suggesting possible poor readings.

Gene 2505169086 is a possible aspartyl/glutamyl-tRNA (Asn/Gln) amidotransferase fragment and/or amidase due to homology and close active site conservation with a verified Asn/Gln-tRNA amidotransferase (EC 6.3.5.-) found in PDB (O66610, 49% identity, E-value 1.18E-49) [[7]](http://198.128.28.118/genome/converter.php#ref-7). Active site mismatches included A200-S, L196-A,V197-Y, and A198-L. These polar residues that are used to stabilize the enzyme reaction are not strictly conserved across other known proteins.

Reverse BLAST searches with a known amidase (Q55424) revealed that it was also homologous with amidase (49% identity, E-value 1.18E-49). Experimentation is necessary to verify the functions of this enzyme.

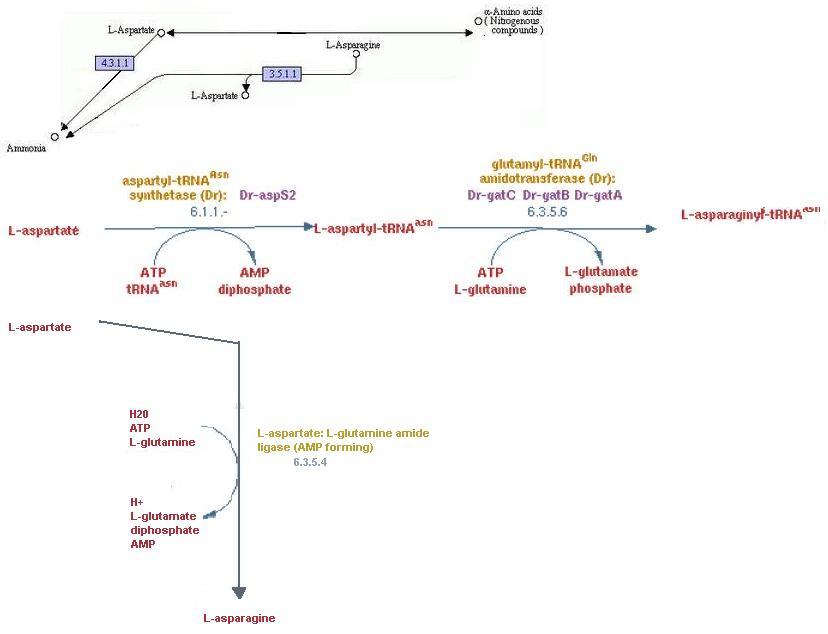
Gene 2505168184 possibly codes for an amidase fragment and/or an amidotranferase (subunit A) fragment due to homology. It possessed weak homology with verified EC 6.3.5.- enzymes found in Swissprot (Q9X0Z9, 43% identity, E-value 1.45E-07) [[4]](http://198.128.28.118/genome/converter.php#ref-4). Reverse BLAST searches with a known amidase (Q55424) revealed that it was also weakly homologous with amidase (39% identity, E-value 2e-4). It was significantly shorter (72 aa) than most amidases and amidotransferases (~500 aa). Experimentation is necessary to verify the functions of this enzyme.

Gene 2505168082 possible codes for an amidase fragment and/or an aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase (subunit A) fragment. It possessed weak homology with verified EC 6.3.5.- enzymes found in Swissprot (P63488, 30% identity, E-value 2.03E-22) [[5]](http://198.128.28.118/genome/converter.php#ref-5). Reverse BLAST searches with a known amidase (Q55424) revealed that it was also weakly homologous with amidase (31% identity, E-value 1e-21). It was significantly shorter (193 aa) than most amidases and amidotransferases (~500 aa). Experimentation is necessary to verify the functions of this enzyme.

**Other Genomes**

IMG/ER predicted several genes in asparagine synthesis and aspartate conversion to be missing. Reverse BLAST searches using enzymes with verified functions (from cyanobacteria if present) were used to find any homologs within the *Microcoleus* genome. The results of this showed that asparagine synthetase, adp-forming and amp-forming, and glutamin-(asparagin-)ase most likely do not exist in *Microcoleus*. (Sample enzyme for EC 6.3.1.1: P00963; 3.5.1.38: P10182; 6.3.1.4 has no known associated enzyme).

**Final Annotation**



S Fig 4. Annotated pathways for Asparagine synthesis

**Functional Studies**

Although Microcoleus probably undergoes L-asparagine tRNA dependent synthesis due to the prediction of the three encoding (Gat) subunits. However, if it is necessary to determine the exact function of those genes predicted to be amidase/amidotransferase fragments, experimentation might include getting a better sequence of the genes region.

Asparagine synthase (glutamine hydrolyzing) function could be tested by constructing mutant strains of *Microcoleus* lacking the possible genes and observing what affect that has on growth in the presence of asparagine, which was done in a study on *Bacillus subtilis*[[8].](http://198.128.28.118/genome/converter.php#ref-8)

**Discussion**

From this gene annotation, it is very probable that *Microcoleus Vaginatus* can synthesize its own Asparagine through various pathways. Although there was some uncertainty concerning genes encoding the tRNA dependent pathway, the rest of the annotation was straightfoward.

**References**

[1] Briand C, Poterszman A, Eiler S, Webster G, Thierry J, Moras D. "An intermediate step in

the recognition of tRNA(Asp) by aspartyl-tRNA synthetase." J Mol Biol. Jun 16, 2000;

299(4):1051-60.

[2] Dieterich DC, Landwehr M, Reissner C, Smalla KH, Richter K, Wolf G, Böckers TM,

Gundelfinger ED, Kreutz MR. "Gliap--a novel untypical L-asparaginase localized to rat

brain astrocytes." J Neurochem. 2003 Jun; 85(5):1117-25.

[3] Hejazi M, Piotukh K, Mattow J, Deutzmann R, Volkmer-Engert R, Lockau W. "Isoaspartyl

dipeptidase activity of plant-type asparaginases." Biochem J. May 15, 2002; 364(Pt

1):129-36.

[4] Ito T, Yokoyama S. "Two enzymes bound to one transfer RNA assume alternative

conformations for consecutive reactions." Nature. Sep 30, 2010; 467(7315):612-6.

[5] Nakamura A, Yao M, Chimnaronk S, Sakai N, Tanaka I. "Ammonia channel couples

glutaminase with transamidase reactions in GatCAB." Science. Jun 30, 2006;

312(5782):1954-8.

[6] Sun DX, Setlow P. "Cloning, nucleotide sequence, and expression of the Bacillus subtilis ans

operon, which codes for L-asparaginase and L-aspartase." J Bacteriol. 1991 Jun;

173(12):3831-45.

[7] Wu J, Bu W, Sheppard K, Kitabatake M, Kwon ST, Söll D, Smith JL. "Insights into

tRNA-dependent amidotransferase evolution and catalysis from the structure of the

Aquifex aeolicus enzyme." J Mol Biol. Aug 28, 2009; 391(4):703-16.

[8] Yoshida K, Fujita Y, Ehrlich SD. "Three asparagine synthetase genes of Bacillus subtilis." J

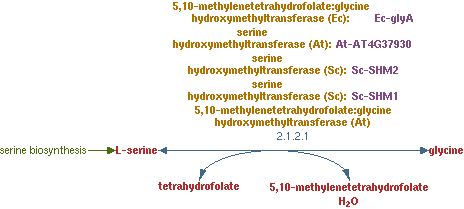
Bacteriol. 1999 Oct; 181(19):6081-91.

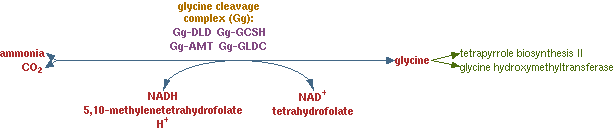
## Glycine

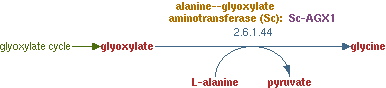
There are 4 different pathways for glycine biosynthesis. When the cells are grown with glucose as the carbon source, two alternative pathways operate. In one pathway L-threonine aldolase, produces glycine from L-threonine (which is produced from the glycolytic intermediate oxaloacetate). In the other pathway glycine is formed from L-serine (a product of 3-phospho-D-glycerate, another glycolytic intermediate). When the cells are grown with a non-fermentable carbon source, such as ethanol and acetate , glycine is produced from glyoxylate, a product of the glyoxylate cycle, by the enzyme alanine, glyoxylate aminotransferase 1. Also, glycine can be biosynthesized by the reverse reactions of the glycine cleavage complex.

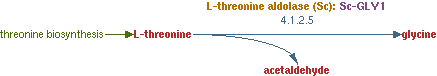
Glycine is used in protein biosynthesis, and glycine is also used in the biosynthesis of other compounds such as tetrapyrroles and purines. Along with serine, glycine is an important donor of one-carbon units via the tetrahydrofolate pathways.

**Automated Annotation**









S Fig. 1: MetaCyc pathway maps of diferent glycine biosynthesis pathways

**Serine hydroxymethyltransferase**

Gene OID: 2505168407 EC: 2.1.2.1

This gene is predicted to code for serine hydroxymethyltransferase. A NCBI BLAST search gave only serine hydroxymethyltransferases as hits [[2]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-2) with e-values of 0 and 62% to 86% identity. A PDB BLAST search gave as a top hit an article on serine hydroxymethyltransferase [[7]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-7) with an e-value of 1.56754E-142 and 61% identity. This article mentioned active sites, and out of these mentioned active sites, 21 out of 22 of them matched. The sole mismatch was R58K. The R58 residue was not mentioned in the article as having specific catalytic funciton, and the R58 and K58 residues are very similar, both having long carbon backbones with nitrogens on them as side groups. Additionally, a reverse BLAST search gave this gene as the top hit, so this gene is strongly predicted to code for serine hydroxymethyltransferase.

**Glycine cleavage system H protein**

Gene OID: 2505168935

This gene is predicted to code for the glycine cleavage system H protein. A NCBI BLAST search gave only other glycine cleavage system H proteins as hits {{NCBI: Q8YNF8}} with e-values from 3E-59 to 3E-32 and 39% to 66% identity. A PDB BLAST search then gave an article on the crystal structure of the H protein of the glycine cleavage system [[5]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-5) with an e-value of 2.16295E-26 and 44% identity. There were 8 active sites mentioned in this 8article, and 7 out of 3 of them matched, with the mismatch being S66E. The article did not mention a specific catalytic function for residue S66. A multiple sequence alignment was done on T-Coffee, using reviewed cyanobacterial FASTA sequences from UniProt, to shed light on this. This alignment showed that at site 66, only one other cyanobacteria had the S66 residue, while almost all the rest had the E66 residue. This, along with the fact that a reverse BLAST search gives the gene as the top hit, leads one to predict that the gene codes for the H protein.

**Alanine-glyoxylate aminotransferase**

Gene OID: 2505170753 EC: 2.6.1.44

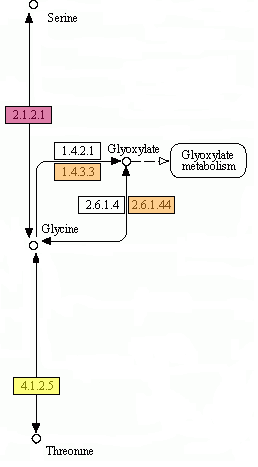
This gene is predicted to code for alanine-glyoxylate aminotransferase. A NCBI BLAST search gave one alanine-glyoxylate aminotransferase as a hit [[6]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-6), with an e-value of 5E-28 and 28% identity. A PDB BLAST search then gave an article on the crystal structure of alanine-glyoxylate aminotransferase as a top hit [[1]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-1) with an e-value of 6.72476E-49 and 35% identity. 8 out of 9 active sites mentioned in the article matched. The only mismatch was a deletion at site 83 of residue T83. The article did not mention T83 as having catalytic function, and it also said that the T83 was not completely conserved. Furthermore, a reverse BLAST search gave this gene as the top hit, so this gene is predicted to code for alanine-glyoxylate aminotransferase.

**Threonine aldolase**

Gene OID: 2505166185 EC: 4.1.2.5

This gene is predicted to code for threonine aldolase. A NCBI BLAST search gave threonine aldolases as top hits [[4]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-4) with e-values from 2E-91 to 5E-05 and 23% to 43% identity. A PDB BLAST search gave as a top hit an article on threonine aldolase crystal structure [[3]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-3) with an e-value of 3.22317E-7 and 25% identity. The active sites could not be checked because the the multiple sequence alignment on PDB didn’t match up. A reverse BLAST search did give this gene as the top hit though. However, the lack of active site confirmation and the weak PDB BLAST hit make this gene one that should be further studied.

**Comparison with Other Genomes**





S Fig. 2: KEGG pathway map for glycine biosynthesis for comparison with other cyanobacteria

The four enzymes needed for the four glycine synthesis pathways were all found in at least some other cyanobacteria. These four enzymes that were found in *Microcoleus vaginatus*, EC 4.1.2.5, 2.1.2.1, 1.4.3.3 and 2.6.1.44, are not unique to the organism but are found in other cyanobacteria as well. Additionally, there were two enzymes predicted to be able to catalyze two of the pathways, EC 1.4.2.1 and 2.6.1.4, but neither of these enzymes were found in cyanobacteria. These enzymes were not found in *Microcoleus vaginatus* either.

From S Figure 2, it seems that the pathway from serine via EC 2.1.2.1 is the most common glycine synthesis pathway among cyanobacteria. The other glycine synthesis pathways seem to be less common among cyanobacteria.

**Functional Studies**

Functional studies should be done on gene 2505166185. This gene is predicted to code for threonine aldolase (EC 4.1.2.5), but the PDB BLAST search gave a weak hit to an article on threonine aldolase structure. Furthermore, this article mentioned active sites that were not able to be compared. An experiment should be done in the lab to see if *Microcoleus vaginatus* expresses threonine aldolase. If it does, this gene is most likely the gene that codes for it. A second experiment should then be done with this gene knocked out, and it should be seen if the enzyme is still expressed.

**Discussion**

*Microcoleus vaginatus* is predicted to be able to synthesize its own glycine. Four different pathways were predicted to be available to *Microcoleus vaginatus* for glycine biosynthesis. Studies should be done to see which other cysteine pathways are favored for the organism, and why. It might be the case that certain pathways are suited for certain environmental conditions. It makes sense for the organism to have multiple pathways for cysteine biosynthesis. This way, it can use different pathways depending on which substrates are abundant. If one substrate is lacking, the organism can just switch to another pathway whose substrate is abundant. Metacyc predicts that

*Microcoleus vaginatus* is known from experiment to be able to synthesize its own glycine, since it experimentally grows without being given any amino acid sources. This annotation report isn’t surprising since it predicts that *Microcoleus vaginatus* can synthesize glycine, as is known to be true in real life.

**References**

[1] Han GW, Schwarzenbacher R, Page R, Jaroszewski L, Abdubek P, Ambing E, Biorac T,

Canaves JM, Chiu HJ, Dai X, Deacon AM, DiDonato M, Elsliger MA, Godzik A, Grittini C,

Grzechnik SK, Hale J, Hampton E, Haugen J, Hornsby M, Klock HE, Koesema E,

Kreusch A, Kuhn P, Lesley SA, Levin I, McMullan D, McPhillips TM, Miller MD, Morse A,

Moy K, Nigoghossian E, Ouyang J, Paulsen J, Quijano K, Reyes R, Sims E, Spraggon

G, Stevens RC, van den Bedem H, Velasquez J, Vincent J, von Delft F, Wang X, West B,

White A, Wolf G, Xu Q, Zagnitko O, Hodgson KO, Wooley J, Wilson IA. "Crystal structure

of an alanine-glyoxylate aminotransferase from Anabaena sp. at 1.70 A resolution

reveals a noncovalently linked PLP cofactor." Proteins. Mar 1, 2005; 58(4):971-5.

[2] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[3] Kielkopf CL, Burley SK. "X-ray structures of threonine aldolase complexes: structural basis of

substrate recognition." Biochemistry. Oct 1, 2002; 41(39):11711-20.

[4] Liu JQ, Ito S, Dairi T, Itoh N, Kataoka M, Shimizu S, Yamada H. "Gene cloning, nucleotide

sequencing, and purification and characterization of the low-specificity L-threonine

aldolase from Pseudomonas sp. strain NCIMB 10558." Appl Environ Microbiol. 1998

Feb; 64(2):549-54.

[5] Nakai T, Ishijima J, Masui R, Kuramitsu S, Kamiya N. "Structure of Thermus thermophilus

HB8 H-protein of the glycine-cleavage system, resolved by a six-dimensional

molecular-replacement method." Acta Crystallogr D Biol Crystallogr. 2003 Sep; 59(Pt

9):1610-8.

[6] Takada Y, Noguchi T. "Characteristics of alanine: glyoxylate aminotransferase from

Saccharomyces cerevisiae, a regulatory enzyme in the glyoxylate pathway of glycine

and serine biosynthesis from tricarboxylic acid-cycle intermediates." Biochem J. Oct 1,

1985; 231(1):157-63.

[7] Trivedi V, Gupta A, Jala VR, Saravanan P, Rao GS, Rao NA, Savithri HS, Subramanya HS.

"Crystal structure of binary and ternary complexes of serine hydroxymethyltransferase

from Bacillus stearothermophilus: insights into the catalytic mechanism." J Biol Chem.

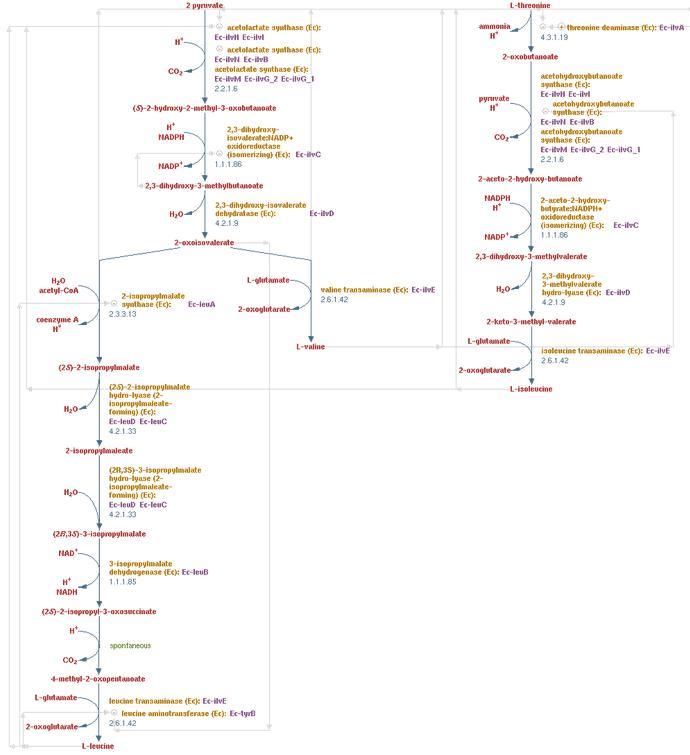
May 10, 2002; 277(19):17161-9.

## Leucine, Valine and Isoleucine

The synthesis of leucine, valine and isoleucine within an organism is not found in organisms. Leucine, valine and isoleucine are essential amino acids for humans for example, since humans cannot synthesize them and must ingest them. *Microcoleus vaginatus* on the other hand, has all the enzymes necessary for leucine, valine and isoleucine biosynthesis, so it can produce these amino acids on its own.

Leucine, valine and isoelucine are each synthesized by their own similar but unique pathways. In each pathway, the final step involves EC 2.6.1.42 converting L-glutamate and another substrate to the corresponding amino acid and 2-oxoglutarate. The substrate that is converted is specific to the amino acid being produced.

**Automated Annotation**



S Fig. 1: Automated MetaCyc K, V, I Biosynthesis Pathway (<http://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=BRANCHED-CHAIN-AA-SYN-PWY&detail-level=2&detail-level=1>)

**3-isopropylmalate dehydrogenase**

Gene OID: 2505169457 EC 1.1.1.85

This gene is predicted to code for the 3-isopropylmalate dehydrogenase enzyme (EC 1.1.1.85). An NCBI BLAST search gave 3-isopropylmalate dehydrogenase enzymes as top hits, with the top hit having an e-value of 0 and a query coverage of 96%. [[6]](http://198.128.31.59/genome/converter.php#ref-6) A PDB BLAST search gave other 3-isopropylmalate dehydrogenase enzymes as top hits, with almost 100% active site conservation, with an e-value of 9.15177E-95 and a percent identity of 52%. [[17]](http://198.128.31.59/genome/converter.php#ref-17) A reverse BLAST search also gave the gene as a top hit, with an e-value of 0.

**Acetolactate synthase**

Gene OID: 2505170041 EC 2.2.1.6

This gene is predicted to code for the acetolactate synthase enzyme (EC 2.2.1.6). An NCBI BLAST search gave acetolactate synthase enzymes as top hits, with the top hit having an e-value of 0 and a query coverage of 92%. [[18]](http://198.128.31.59/genome/converter.php#ref-18) A PDB BLAST search initially gave acetolactate synthase enzyme structures within eukaryotes as top hits. But once the search was refined to only include structures within bacteria, the PDB BLAST search did not give any hits to EC 2.2.1.6. A reverse BLAST search, however, did give the gene as a top hit with an e-value of 0. Although the PDB results were inconclusive, it is very likely from the NCBI BLAST hits that the gene 2505170041 codes for EC 2.2.1.6.

Gene OID: 2505170540 EC 2.2.1.6

This gene might code for the acetolactate synthase enzyme (EC 2.2.1.6). An NCBI BLAST search gave acetolactate synthase enzymes as top hits, with the top hit having an e-value of 1e-110 and a query coverage of 96%. [[12]](http://198.128.31.59/genome/converter.php#ref-12) The PDB BLAST search did not give any hits to EC 2.2.1.6. Most of the hits were of the 2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase enzyme structure (EC 2.2.1.9). According to IMG/ER, the gene 2505167327 in *Microcoleus vaginatus* codes for EC 2.2.1.9. Also, EC 2.2.1.9 within *Microcoleus vaginatus* is predicted to be a part of the 1,4-dihydroxy-2-naphthoate biosynthesis I pathway, according to Metacyc. A reverse BLAST search, however, did give gene 2505170540 as a top hit with an e-value of 9e-99.

**Dihydroxy-acid dehydratase**

Gene OID: 2505171024 EC: 4.2.1.9

Gene is predicted to code for dihydroxy-acid dehydratase. A NCBI BLAST search gave as top hits other dihydroxy-acid dehydratases [[5]](http://198.128.31.59/genome/converter.php#ref-5) with e-values of 0 and around 80% identity. A reverse BLAST search gave gene 2505171024 as the top hit. A PDB BLAST search was also done, but no hits with EC 4.2.1.9. The PDB database doesn’t have any articles tagged with EC 4.2.1.9, so the PDB search was inconclusive. However, from the NCBI BLAST hits, it is very likely that the gene codes for EC 4.2.1.9.

**Threonine ammonia-lyase, biosynthetic, long form**

Gene OID: 2505166192 EC: 4.3.1.19

Gene is predicted to code for EC 4.3.1.19. A NCBI BLAST search gave as top hits threonine deaminases (same as threonine ammonia-lyase) [[9]](http://198.128.31.59/genome/converter.php#ref-9) with e-values of around 0 and around 50% identity.

A PDB BLAST search gave as a top hit an article on theronine deaminase [[15]](http://198.128.31.59/genome/converter.php#ref-15) with an e-value of 9.9479E-51 and 37% identity. This article mentioned active sites for the protein. There were active sites involved in binding with two different cofactors: PLP and CMP. For binding with PLP, 15 out of 19 active site residues matched. The mismatches were I59L, F63Y, P182A and S311C. The article did not mention anything special about sites 59, 63 or 182. For site 59, isoleucine and leucine are similar in structure and properties, only having different branching. For site 63, both phenylalanine and tyrosine are aromatic, with the only difference being a phenol group versus a benzene ring. For site 182, proline and alanine are not similar in structure, but there is no mention that proline has specific function. For site 311, serine is mentioned to hydrogen bond with its carboxylic acid oxygen, which cysteine also has. It is also mentioned that serine is important because it is a neutral amino acid, and it is mentioned that serine, cysteine or threonine can be present at site 311.

For binding with CMP, only 3 out of 13 active site residues matched. This suggests that the protein cannot bind CMP. According to the article, CMP is not necessary for the protein to function, but it does help the protein to catalyze its substrate more efficiently. Therefore, this gene is still predicted to code for EC 4.3.1.19.

Gene OID: 2505167467 EC 4.3.1.19

A NCBI BLAST search gave as top hits serine racemases [[20]](http://198.128.31.59/genome/converter.php#ref-20) with low e-values (2E-110 to 4E-73) and high percent identity (45% to 50%). NCBI also gave hits to threonine dehydratases [[14]](http://198.128.31.59/genome/converter.php#ref-14) with e-values from 9E-71 and 4E-62 and around 40% identity.

A PDB BLAST search gave as one of the top hits an article on serine racemase [[4]](http://198.128.31.59/genome/converter.php#ref-4) with an e-value of 6.19206E-93 and 53% identity. Active sites were mentioned in the article. 17 out of 20 of the mentioned active sites matched. The three mismatches were L182C, G212A and Q238R. None of the sites were mentioned in the article to have special function, so it seems that the active site matching is good. This would suggest that the gene could code for serine racemase.

However, there was another PDB BLAST hit to an article on threonine deaminase [[15]](http://198.128.31.59/genome/converter.php#ref-15), with an e-value of 1.21689E-64 and 43% identity. This was the same article as for the previous gene. There were active sites mentioned for binding PLP and CMP. For the PLP binding site, 15 out of 19 active sites matched. The mismatches were I59F, F63C, P182C and I189L. Sites 59, 63, 182 and 189 were not mentioned in the article as being special, so it seems like the active site matching is good. For the CMP binding site, 9 out of 13 sites matched, with the mismatches being N34E, Y35T, D119G and Q275E. The article says N34 is important due to its side chain nitrogen, which E34 has none of. For site 35, there is no mention Y35 being important, so it is not unusual that it is changed to T. For site D119, the article mentioned its side-chain oxygen is needed for hydrogen bonding. G119 doesn’t have a side-chain oxygen. For site Q275, the article mentioned the side-chain oxygen as being important for hydrogen bonding. E275 has a side-chain oxygen as well. It seems that the PLP binding site exists, but whether or not the CMP binding site exists is more unclear.

This gene could code for serine racemase or threonine deaminase. The hits to serine racemase are better than for theronine deaminase, but further research needs to be done to see which protein the gene codes for, or to see if the gene codes for a bifunctional protein.

**Branched-chain amino acid aminotransferase**

Gene OID: 2505166246 EC 2.6.1.42

Gene was predicted to code for EC 2.6.1.42. A NCBI BLAST search gave other branched-chain amino acid aminotransferases [[3]](http://198.128.31.59/genome/converter.php#ref-3) as top hits, with e-values of 6E-74 to 5E-18 and around 35% identity. A PDB BLAST search gave an article on branched-chain amino acid aminotransferase as the top hit [[10]](http://198.128.31.59/genome/converter.php#ref-10), with an e-value of 1.81976E-43 and 33% identity. Active site residues were mentioned in the article, and 9 out of 11 of these matched. The mismatches were F36L and V109-. The article mentioned F36 as being important for forming part of a hydrophobic core. L36 would be able to help form a hydrophobic core as well, since it is also hydrophobic. V109 was not mentioned to have specific function in the article. V109 was a deletion; there was no amino acid from the query sequence that lined up to V109 of the PDB sequence. To double-check both of these results, a multiple sequence-alignment was done on T-Coffee, using cyanobacteria FASTA sequences from UniProt as a reference. This showed that at site 36, the F36 residue was completely conserved among all other cyanobacteria. This could be a mutation specific to *Microcoleus vaginatus*. The alignment also forced an alignment at site 109, and showed the corresponding amino acid to be P109. This P109 residue was almost completely conserved among the other cyanobacteria, with only one cyanobacteria not having a P109 residue. Additionally, the multiple sequence alignment showed good overall matching between this gene and other cyanobacterial genes for EC 2.6.1.42. Despite the F36L mismatch, the rest of the data still suggests that this gene codes for EC 2.6.1.42.

**Ketol-acid reductoisomerase**

Gene OID: 2505169860 EC: 1.1.1.86

This gene is predicted to code for Ketol-acid reductoisomerase due to strong homology with experimentally verified enzymes with EC 1.1.1.86 and 100% active site conservation. Strong homologs were found in PDB, which supplied the active site residues, (Uniprot: P05793, 35% identity, E-value 2.20E-29) [[19]](http://198.128.31.59/genome/converter.php#ref-19) and Swissprot databases (Swissprot: P29107, 86% identity, E-value 0) [[7]](http://198.128.31.59/genome/converter.php#ref-7).

**2-isopropylmalate synthase/homocitrate synthase family protein**

Gene OID: 2505167522 EC: 2.3.3.13/2.3.3.14

This gene is predicted to code for a protein with EC 2.3.3.13 (2-isopropylmalate synthase) and/or EC 2.3.3.14 (Homocitrate synthase) functions. This is inferred from the strong homology with experimentally verified enzymes of both functions in PDB and from active site conservation. (For EC:2.3.3.13: Uniprot: P96420, 25% identity, E-value 3.84E-12 [[8]](http://198.128.31.59/genome/converter.php#ref-8); for EC: 2.3.3.14: Uniprot: O87198, 30% identity, E-value 1.12E-29) [[11]](http://198.128.31.59/genome/converter.php#ref-11). Literature [[16]](http://198.128.31.59/genome/converter.php#ref-16) and a multiple sequence alignment with T-Coffee suggests that the genes coding for these functions are very similar. 100% active residue conservation with homocitrate and close conservation with 2-isoproplymalate synthase (mismatches are E218-N and T254-E) supports this. If necessary further lab work should be done to make a distinction.

**2-isopropylmalate synthase**

Gene OID: 2505165612 EC: 2.3.3.13

This gene is predicted to code for 2-isopropylmalate synthase due to strong homology with experimentally verified enzymes with EC 2.3.3.13 and 100% active site conservation. Strong homologs were found in PDB with EC numbers 2.3.3.13 and 2.3.3.14, but the EC 2.3.3.13 homologs active sites were conserved 100% while 2.3.3.14 active sites were not conserved. Homologs were found via PDB (Uniprot: P96420, 30% identity, E-value 3.12E-22) [[1]](http://198.128.31.59/genome/converter.php#ref-1) and Swissprot (Uniprot: P48575, 83% identity, E-value 0) [[16]](http://198.128.31.59/genome/converter.php#ref-16).

However, given that homocitrate synthase is very similar to 2-isoproplymalate (see annotation of Gee 2505167522) further research may be needed to make a more clear distinction.

**3-isopropylmalate/(R)-2-methylmalate dehydratase, large subunit**

Gene OID: 2505166495 EC: 4.2.1.33/35

This gene is predicted to code for the large subunit bifunctional protein Homoaconitate hydratase family protein/3-isopropylmalate dehydratase due to strong homology with experimental enzymes with EC 4.2.1.33 and 4.2.1.35. A homolog was found in Swissprot (Uniprot: P81291, 45% identity, E-value: 3.00E-127) [[2]](http://198.128.31.59/genome/converter.php#ref-2) but unfortunately PDB BLAST only yielded mitochondrial aconitase, which is a close relative but not a very strong homolog to other large subunit dehydtratase proteins found in UniProt.

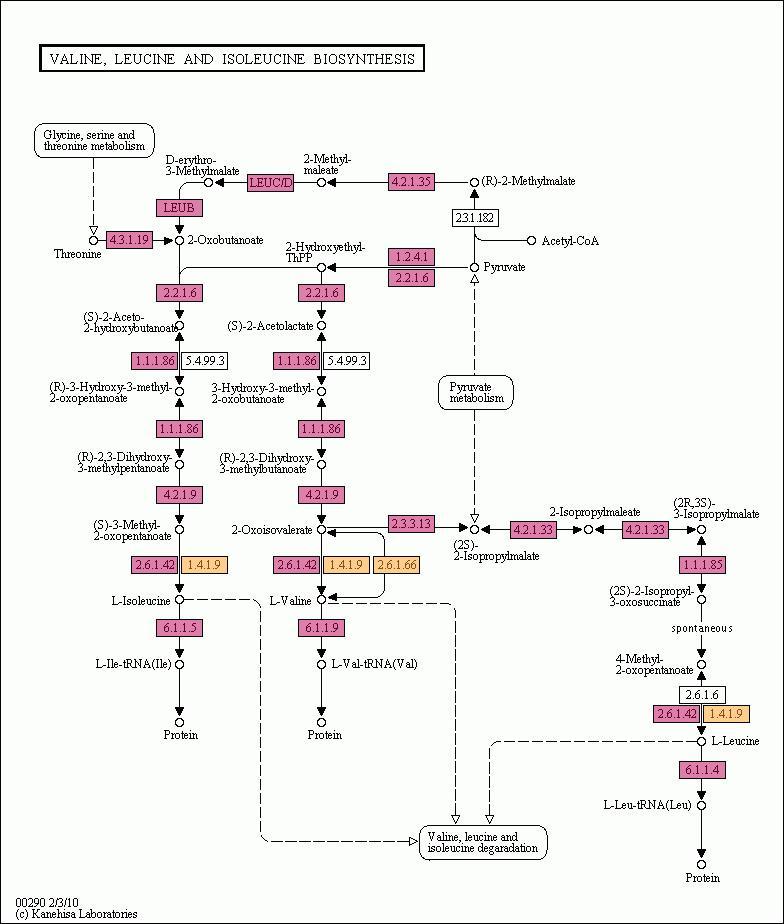
Because no active site residues were found and the protein is a close relative of aconitate hydratase 1 and homoaconitase, further investigation may be required to conclusively identify the function of this gene.

**3-isopropylmalate/(R)-2-methylmalate dehydratase, small subunit**

Gene OID: 2505167112 EC: 4.1.2.33/35

This gene is predicted to code for the small subunit bifunctional protein 3-isopropylmalate dehydratase due to strong homology with experimentally enzymes with EC 4.2.1.33 and 4.2.1.35 and 100% active site conservation. Homologs were found in PDB (Uniprot: O59393, 48% identity, E-value 7.94E-11) [[21]](http://198.128.31.59/genome/converter.php#ref-21) and Swissprot databases (Uniprot: Q30RK1, 41% identity, E-value 1e-31) [[13]](http://198.128.31.59/genome/converter.php#ref-13).

**Comparison with other Genomes**





S Fig. 2: Comparison with other cyanobacterial genomes in IMG/ER

S Figure 2 was taken from KEGG and is not exactly the same as S Figure 1, which was taken from MetaCyc. Following the parts of the pathway that are the same as the MetaCyc one shows that all the genes annotated in the report were found in more than 75% of other cyanobacteria. The genes were not unique to *Microcoleus vaginatus*. These results suggest that cyaonbacteria in general are able to undergo leucine, valine and isoleucine biosynthesis.

**Functional Studies**

Functional studies should be done on genes 2505166192 and 2505167467. Gene 2505166192 was predicted to be a threonine ammonia-lyase, but active site matching showed that it could not bind CMP, which according to the PDB article helps the enzyme to catalyze its substrate more efficiently. Studies should be done to see if this is the case, and if the enzyme can still function without binding CMP.

Gene 2505167467 was predicted to be a threonine ammonia-lyase, but the top NCBI and PDB BLAST hits for this gene were to serine racemases. An active site residue comparison showed good matching, so this gene could code for a serine racemase. However, there were also hits to threonine ammonia-lyases, and active site matching showed that this gene coded for a protein that was more probable to be able to bind CMP. An experiment needs to be done to see which protein the gene actually codes for, because from the annotation it is unclear.

Gene 2505166246 should also be checked. This gene is predicted to code for a branched-chain amino acid aminotransferase. There was one problematic active site mismatch though, F36L. The F36 residue was conserved among other cyanobacteria, so it is unusual that the residue is L36 for the gene. Functional studies should be done to double check that the gene does indeed code for branched-chain amino acid aminotransferase. If it does, the L36 residue is interesting because it does not appear in other cyanobacteria.

Genes 2505170041 and 2505170540 should also be checked. These genes were both predicted to code for EC 2.2.1.6, so it should be checked if indeed they both code for proteins with the same function, or if only one of them codes for a protein with function for EC 2.2.1.6. It should also be checked to see if either of these genes actually code for EC 2.2.1.6, because no PDB hits with EC 2.2.1.6 were found from BLAST searches, although this might be because PDB doesn’t have any articles with tags to bacterial EC 2.2.1.6 in its database yet. It should also be seen if 2505170540 is a misannotated gene and if it actually has function for EC 2.2.1.9, because most of the PDB BLAST hits are to structures with tags to EC 2.2.1.9. It would be interesting if this is the case, because there is already a gene that codes for EC 2.2.1.9, and there would be redundancy there.

Gene 2505166495 should also be checked. This gene was predicted to code for a protein with function for EC 4.2.1.33 and 4.2.1.35. However, PDB BLAST searches were not conclusive, so further functional studies should be done to see if this gene codes for a bifunctional protein or a protein with function for only one of the EC numbers.

**Discussion**

There were hits for all the genes necessary for leucine, valine and isoleucine biosynthesis, but for some of the hits, it was unclear if the genes actually had the predicted function. For gene 2505166192, it is predicted to be a threonine ammonia-lyase, but without the ability to bind CMP, which is stated in the literature to be an important cofactor. Studies should be done to see if there are any advantages to not binding CMP, because the literature implies that not binding CMP is not advantageou in any way. Studies should also be done to see if other organisms have threonine ammonia-lyases that do not bind CMP.

For gene 2505167467, further studies need to be done to determine if the the gene codes for serine racemase, threonine ammonia-lyase or a bifunctional protein. IF this gene codes for threonine ammonia-lyase, this version seems to be able to bind CMP. Studies should be done to see why there are two genes for two different variations of theronine ammonia-lyase: one that can’t bind CMP and one that can. It should be seen if there is some evolutionary advantage to being able to or not being to bind CMP, and if other organisms are similar in this way.

For gene 2505166246, if this gene is confirmed to code for branched-chain amino acid aminotransferase, it is interesting that it has residue L36. Studies should be done to see if this was a beneficial mutation, and if this mutation is found in other organism or why it is not found in other organsims.

For genes 2505170041 and 2505170540, they were both predicted to code for EC 2.2.1.6. Assuming both these genes actually do code for EC 2.2.1.6, it should be seen if there is an evolutionary advantage to having two genes that code for the same protein, or if the genes are activated under different external conditions to code for their respective proteins. If gene 2505170540 is misannotated, it should be seen why there are two genes coding for EC 2.2.1.9, and similar questions should be asked as for EC 2.2.1.6.

For gene 2505166495, studies should be done to see why the gene codes for a bifunctional protein. Studies could be done with another organism that has two separate proteins instead of a bifunctional protein, and comparisons could be made between the two organisms. It should also be seen if this bifunctional protein can be found in other organisms, and what other kinds of organisms, and it could be observed if they are similar or different compared to *Microcoleus vaginatus*.

**References**

[1] Chanchaem W, Palittapongarnpim P. "A variable number of tandem repeats result in

polymorphic alpha -isopropylmalate synthase in Mycobacterium tuberculosis."

Tuberculosis (Edinb). 2002; 82(1):1-6.

[2] Drevland RM, Waheed A, Graham DE. "Enzymology and evolution of the pyruvate pathway to

2-oxobutyrate in Methanocaldococcus jannaschii." J Bacteriol. 2007 Jun;

189(12):4391-400.

[3] Feild MJ, Nguyen DC, Armstrong FB. "Amino acid sequence of Salmonella typhimurium

branched-chain amino acid aminotransferase." Biochemistry. Jun 13, 1989;

28(12):5306-10.

[4] Goto M, Yamauchi T, Kamiya N, Miyahara I, Yoshimura T, Mihara H, Kurihara T, Hirotsu K,

Esaki N. "Crystal structure of a homolog of mammalian serine racemase from

Schizosaccharomyces pombe." J Biol Chem. Sep 18, 2009; 284(38):25944-52.

[5] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[6] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[7] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo

K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M,

Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[8] Koon N, Squire CJ, Baker EN. "Crystal structure of LeuA from Mycobacterium tuberculosis, a

key enzyme in leucine biosynthesis." Proc Natl Acad Sci U S A. Jun 1, 2004;

101(22):8295-300.

[9] May BJ, Zhang Q, Li LL, Paustian ML, Whittam TS, Kapur V. "Complete genomic sequence of

Pasteurella multocida, Pm70." Proc Natl Acad Sci U S A. Mar 13, 2001; 98(6):3460-5.

[10] Okada K, Hirotsu K, Sato M, Hayashi H, Kagamiyama H. "Three-dimensional structure of

Escherichia coli branched-chain amino acid aminotransferase at 2.5 A resolution." J

Biochem. 1997 Apr; 121(4):637-41.

[11] Okada T, Tomita T, Wulandari AP, Kuzuyama T, Nishiyama M. "Mechanism of substrate

recognition and insight into feedback inhibition of homocitrate synthase from Thermus

thermophilus." J Biol Chem. Feb 5, 2010; 285(6):4195-205.

[12] Peng HL, Wang PY, Wu CM, Hwang DC, Chang HY. "Cloning, sequencing and

heterologous expression of a Klebsiella pneumoniae gene encoding an FAD-independent

acetolactate synthase." Gene. Aug 1, 1992; 117(1):125-30.

[13] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[14] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[15] Simanshu DK, Savithri HS, Murthy MR. "Crystal structures of Salmonella typhimurium

biodegradative threonine deaminase and its complex with CMP provide structural insights

into ligand-induced oligomerization and enzyme activation." J Biol Chem. Dec 22, 2006;

281(51):39630-41.

[16] Stricker O, Masepohl B, Klipp W, Böhme H. "Identification and characterization of the

nifV-nifZ-nifT gene region from the filamentous cyanobacterium Anabaena sp. strain PCC

7120." J Bacteriol. 1997 May; 179(9):2930-7.

[17] Tsuchiya D, Sekiguchi T, Takenaka A. "Crystal structure of 3-isopropylmalate

dehydrogenase from the moderate facultative thermophile, Bacillus coagulans: two

strategies for thermostabilization of protein structures." J Biochem. 1997 Dec;

122(6):1092-104.

[18] Tsuchiya D, Sekiguchi T, Takenaka A. "Crystal structure of 3-isopropylmalate

dehydrogenase from the moderate facultative thermophile, Bacillus coagulans: two

strategies for thermostabilization of protein structures." J Biochem. 1997 Dec;

122(6):1092-104.

[19] Wek RC, Hatfield GW. "Nucleotide sequence and in vivo expression of the ilvY and ilvC

genes in Escherichia coli K12. Transcription from divergent overlapping promoters." J

Biol Chem. Feb 15, 1986; 261(5):2441-50.

[20] Yamauchi T, Goto M, Wu HY, Uo T, Yoshimura T, Mihara H, Kurihara T, Miyahara I, Hirotsu

K, Esaki N. "Serine racemase with catalytically active lysinoalanyl residue." J Biochem.

2009 Apr; 145(4):421-4.

[21] Yasutake Y, Yao M, Sakai N, Kirita T, Tanaka I. "Crystal structure of the Pyrococcus

horikoshii isopropylmalate isomerase small subunit provides insight into the dual

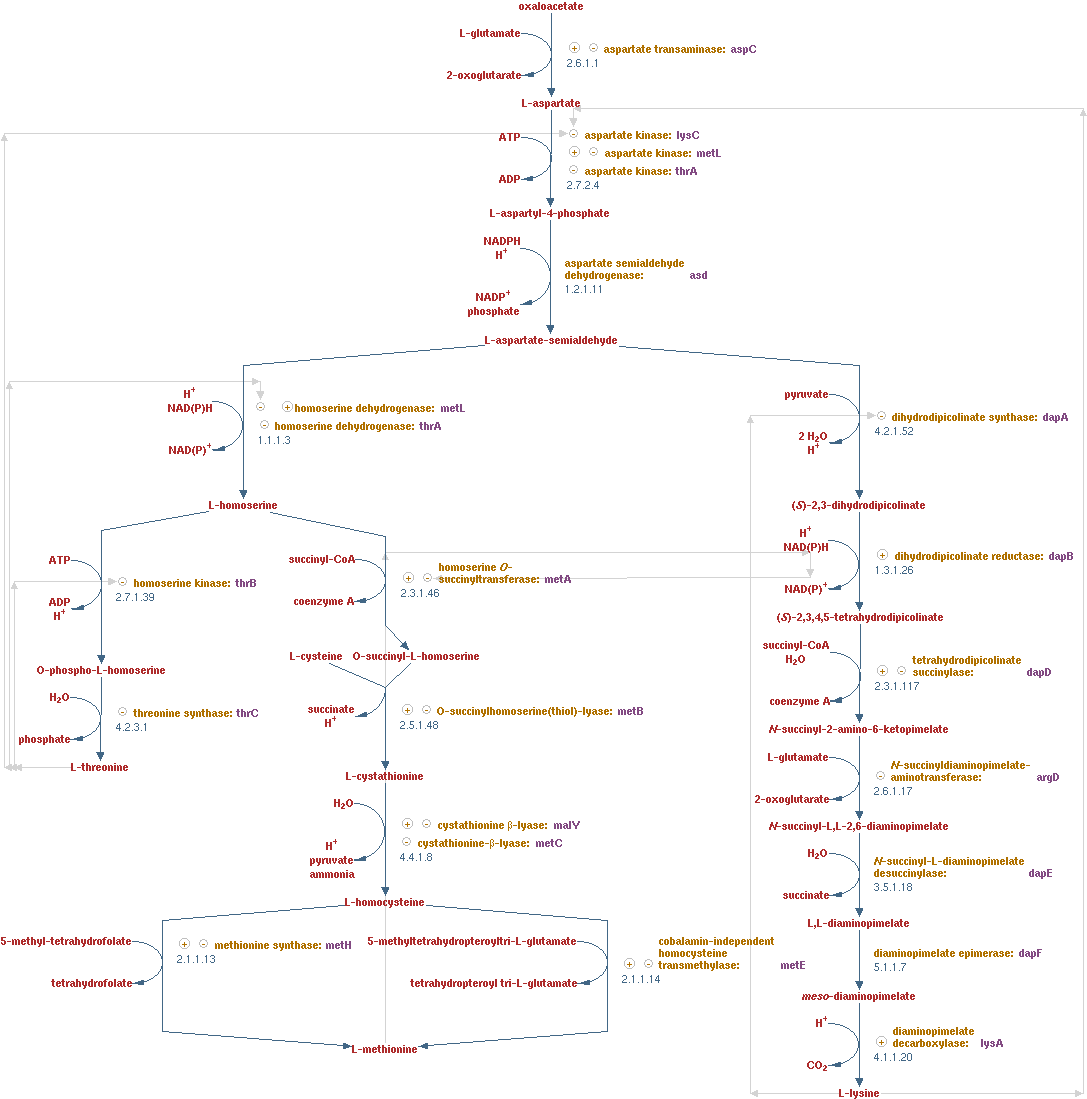
substrate specificity of the enzyme." J Mol Biol. Nov 19, 2004; 344(2):325-33.

## Lysine, Threonine amd Methionine

Lysine, threonine and methionine are 3 of many essential amino acids. Humans cannot synthesize these amino acids, but some other organisms can. An annotation of this pathway was done to see if *Microcoleus vaginatus* was such an organism that could synthesize its own lysine, threonine or methionine.

In this pathway, oxaloacetate is converted to L-aspartate-semialdehyde. This substrate can then be broken down in two pathways. One pathway leads to L-lysine, while the other pathway leads to L-homoserine. L-homoserine can then either by converted to L-threonine or L-methionine via two separate pathways. From the annotation report, it is predicted that *Microcoleus vaginatus* can synthesize its own L-threonine, but not L-lysine or L-methionine.

**Automated Annotation**



S Fig. 1: MetaCyc pathway map of K,T,M Amino Acid Biosynthesis

**Aspartate transaminase**

Gene OID’s: 2505169488 and 2505171077 EC: 2.6.1.1

Two genes (2505169488 and 2505171077) were predicted to code for this enzyme. NCBI BLAST searches for each of the genes [[14]](http://198.128.28.118/genome/converter.php#ref-14) [[7]](http://198.128.28.118/genome/converter.php#ref-7) against the SwissProt database gave other Aspartate transaminase enzymes as top hits. PDB BLAST searches for each of the genes [[16]](http://198.128.28.118/genome/converter.php#ref-16) [[25]](http://198.128.28.118/genome/converter.php#ref-25) gave articles on Aspartate transanimase, and for both of the articles gave active site residues; all these residues matched for both of the genes. An IMG BLAST Search also provided reliable numerical data that both the genes were homologs. The genes were found to be homologs to each other, with 25.14 percent identity and an e value of 1e-25.

**Aspartate Kinase**

Gene OID: 2505166620 EC: 2.7.2.4

This gene is predicted to code for aspartate kinase. A NCBI BLAST search gave as top hits other aspartate kinases [[24]](http://198.128.28.118/genome/converter.php#ref-24) with e-values from 4E-125 to 9E-13 and around 35% to 50% identity. A PDB BLAST search gave an article on aspartate kinase [[21]](http://198.128.28.118/genome/converter.php#ref-21) as a top hit, with an e-value of 0 and 68% identity. Out of the active sites mentioned in the article, 23 out of 25 of them matched. The two mismatches were I299V and I371V. These sites were not mentioned in the article to have specific function, and isoleucine and valine are very similar in structure. Additionally, a reverse BLAST search gave this gene as the top hit. Therefore, the gene is strongly predicted to code for aspartate kinase.

**Aspartate-semialdehyde dehydrogenase**

Gene OID: 2505165750 EC: 1.2.1.11

This gene is predicted to code for aspartate-semialdehyde dehydrogenase. A NCBI BLAST search gave as top hits other aspartate-semialdehyde dehydrogenases [[9]](http://198.128.28.118/genome/converter.php#ref-9) with e-values ranging from 5E-154 to 1E-13, and with anywhere from 40% to 70% identity. A PDB BLAST search gave an article on aspartate-semialdehyde dehydrogenase [[26]](http://198.128.28.118/genome/converter.php#ref-26) with an e-value of 2.17943E-83 and 50% identity. Out of the active sites mentioned in the article, all of them matched (9 out of 9), and a reverse BLAST search gave the same gene as the top hit. This gene is then strongly predicted to code for aspartate-semialdehyde dehydrogenase.

**Dihydrodipicolinate synthase**

Gene OID: 2505165749 EC: 4.2.1.52

This gene is predicted t ocode for dihydrodipicolinate synthase. A NCBI BLAST search gave only dihydrodipicolinate synthases as hits [[12]](http://198.128.28.118/genome/converter.php#ref-12) with e-values from 2E-155 to 3E-74 and around 45% to 70% identity. A PDB BLAST search gave as a hit an article on dihydrodipicolinate synthase crystal structure [[27]](http://198.128.28.118/genome/converter.php#ref-27) with an e-value of 2.92872E-73 and 47% identity. The article mentioned active sites, and out of these active sites, 5 out of 5 of them matched. Additionally, a reverse BLAST search gave the gene as the top hit, so the gene is strongly predicted to code for dihydrodipicolinate synthase.

**Dihydrodipicolinate reductase**

Gene OID: 2505165781 EC: 1.3.1.26

This gene is predicted to code for dihydrodipicolinate reductase. A NCBI BLAST search gave only dihydrodipicolinate reductases as hits [[15]](http://198.128.28.118/genome/converter.php#ref-15) with e-values ranging from a low of 5E-162 to a high of 3E-56 and anywhere from 40% to 80% identity. A PDB BLAST search gave as a top hit an article on dihydrodipicolinate reductase [[3]](http://198.128.28.118/genome/converter.php#ref-3) with an e-value of 1.39284E-47 and 41% identity. Out of the active sites mentioned in the article, 15 out of 19 of them matched up. The mismatches were K9A, V12M, A102I and R155G. In the article, a K9A mutant was made, and it can still function although it has decreased NADPH specificity. V12, A102 and V155 were not mentioned in the article to have specific purpose for the enzyme. A multiple sequence alignment of reviewed cyanobacterial FASTA sequences from UniProt was done to shed more light into the last 3 mismatches. At site 12, no other cyanobacteria had the V12 residue, but rather, the M12 residue was completely conserved. At site 102, three other cyanobacteria had the A102 residue, but more than 70% had the I102 residue. At site 155, only one other cyanobacteria had the R155 residue, but more than 90% had the G155 residue. Also, a reverse BLAST gives this gene as the top hit. It is therefore predicted that the gene codes for dihydrodipicolinate reductase.

**Tetrahydrodipicolinate succinylase**

EC: 2.3.1.117

There is no gene predicted by IMG/ER to code for this protein. A search was done on the UniProt database for cyanobacterial tetrahydrodipicolinate succinylase FASTA sequences, but none were found. Using the FASTA sequences from other bacteria instead, a reverse BLAST was done but gave no hits.

**Succinyldiaminopimelate transaminase**

Gene OID: 2505167581 EC: 2.6.1.17

This gene is predicted to code for succinyldiaminopimelate transaminase. A NCBI BLAST search gave LL-diaminopimelate aminotransferases as hits [[22]](http://198.128.28.118/genome/converter.php#ref-22) with e-values ranging from 2E-109 to 4E-15 and anywhere from 25% to 40% identity. A PDB BLAST search gave as a hit an article on N-succinyldiaminopimelate aminotransferase [[28]](http://198.128.28.118/genome/converter.php#ref-28) with an e-value of 5.3548E-28 and 28% identity. Out of the active sites mentioned, 7 out of 11 of them matched, with the mismatches being A94S, T95Q, S229E and K240R. T95 was mentioned as being important for its hydrogen bonding, which Q95 can’t participate in because it has no hydroxyl group. K240 was mentioned as important for its nitrogen for bonding; R240 also has nitrogen on its side chain that can be used for bonding. Sites 94 and 229 were not mentioned as having specific activity in the paper. A multiple sequence alignment was done on T-Coffee as a double check. The sequences used for comparison were cyanobacterial ones from UniProt. At site 94, no cyanobacteria had the A94 residue; rather, they all had the S94 residue. Similarly, at site 95, no cyanobacteria had the T95 residue, but all of them had the Q95 residue. At site 229, no cyanobacteria had the S229 residue, but all of them had the E229 residue. At site 240, all the other cyanobacteria had the R240 residue. Also, a reverse BLAST search gave this gene as the top hit, so the gene is predicted to code for the protein.

***N*-succinyl-L-diaminopimelate desuccinylase**

EC: 3.5.1.18

No gene was predicted to code for this enzyme by IMG/ER. A search was done on UniProt for cyanobacterial N-succinyl-L-diaminopimelate desuccinylase sequences, but none were found. Instead, bacterial sequences were used and reverse BLASTed against the *Microcoleus vaginatus* genome in IMG/ER. This gave no hits. The fact that no other cyanobacteria were found in UniProt with this protein makes it seem likely that *Microcoleus vaginatus* does not have a gene coding for this protein.

**Diaminopimelate epimerase**

Gene OID: 2505169942 EC: 5.1.1.7

This gene is predicted to code for diaminopimelate epimerase. A NCBI BLAST search gave only diaminopimelate epimerases as top hits {{NCBI: Q8YVD0}} with e-values from 4E-165 to 6E-69 and 42% to 77% identity. A PDB BLAST search gave as a top hit an article on diaminopimelate epimerase [[19]](http://198.128.28.118/genome/converter.php#ref-19) with an e-value of 2.88481E-101 and 62% identity. Out of the active sites mentioned in the article, 17 out of 17 of them matched, and a reverse BLAST search gave this gene as the top hit. This gene is strongly predicted to code for diaminopimelate epimerase.

**Diaminopimelate decarboxylase**

There were two genes predicted by IMG/ER to code for this protein. Gene 2505168421 is predicted to possibly code for the protein, while gene 2505169231 is strongly predicted to code for the protein.

Gene OID: 2505168421 EC: 4.1.1.20

This gene is predicted to code for diaminopimelate decarboxylase with some uncertainty. A NCBI BLAST search gave diaminopimelate decarboxylases as the top hits [[23]](http://198.128.28.118/genome/converter.php#ref-23) with e-values from 2E-57 to 4E-16 and around 25% to 30% identity. A PDB BLAST search gave as a top hit an article on diaminopimelate decarboxylase crystal structure [[20]](http://198.128.28.118/genome/converter.php#ref-20) with an e-value of 1.00181E-29 and 26% identity. Active sites were mentioned in this article, but only 2 out of 6 of them matched. The mismatches were H214Y, Y337H, C362T and E363Y. H214 was important for its hydrogen bonding, which Y214 can also provide. Also, H214 and Y214 are similar in that they both have aromatic rings. Y337 is also important for hydrogen bonding, which H337 could also provide, and as previously mentioned, they are both aromatic. The article does not mention specific purposes for C362 or E363. A reverse BLAST did not give this gene as the top hit, but rather gene 2505169231 (the gene annotated after this one). From these data, it is concluded that the gene could code for diaminopimelate decarboxylase, but that further research needs to be done to confirm this.

Gene OID: 2505169231 EC: 4.1.1.20

This gene is predicted to code for diaminopimelate decarboxylase. A NCBI BLAST search gave other diaminopimelate decarboxylases as top hits [[29]](http://198.128.28.118/genome/converter.php#ref-29) with a range of e-values from 0 to 1E-27 and 26% to 66% identity. A PDB BLAST search gave as a top hit an article on diaminopimelate decarboxylase [[10]](http://198.128.28.118/genome/converter.php#ref-10) with an e-value of 1.59534E-74 and 39% identity. Out of the active sites mentioned in the article, 14 out of 14 of them matched. Additionally, a reverse BLAST search gave the same gene as the top hit; therefore, the gene is strongly predicted to code for diaminopimelate decarboxylase.

**Homoserine dehydrogenase**

Gene OID: 2505171063 EC: 1.1.1.3

This gene is predicted to code for homoserine dehydrogenase. A NCBI BLAST search gave homoserine dehydrogenases as top hits [[4]](http://198.128.28.118/genome/converter.php#ref-4) with e-values from 0 to 1E-29 and anywhere from 28% to 69% identity. A PDB BLAST search gave an article on homoserine dehydrogenase as a top hit [[6]](http://198.128.28.118/genome/converter.php#ref-6) with an e-value of 1.93413E-16 and 29% identity. Out of the active sites mentioned in the paper, 9 out of 11 of them matched. The mismatches were A148G and L150I. The paper did not mention A148 or L150 as having side groups essential to catalytic function. Besides, A and G are very similar in structure, as are L and I. Finally, a reverse BLAST search gives this gene as the first hit, so the gene is predicted to code for homoserine dehydrogenase.

**Homoserine kinase**

Gene OID: 2505169061 EC: 2.7.1.39

This gene is predicted to code for homoserine kinase. A NCBI BLAST search gave only homoserine kinases as hits [[18]](http://198.128.28.118/genome/converter.php#ref-18) with e-values from 4E-138 to 4E-29, and 28% to 58% identity. A PDB BLAST search gave an article on homoserine kinase as a top hit [[30]](http://198.128.28.118/genome/converter.php#ref-30) with an e-value of 2.18329E-9 and 23% identity. Out of the active sites mentioned, only 6 out of 14 of them matched. However, a reverse BLAST search gave this gene as the top hit. It seems that the gene codes for homoserine kinase, but further research should be done because the bad active site matching throws doubt on that conclusion.

**Threonine synthase**

There are two genes predicted by IMG/ER to code for this protein. The first gene, 2505165746, could possibly code for threonine synthase. The second gene, 2505166218, is strongly predicted to code for threonine synthase.

Gene OID: 2505165746 EC: 4.2.3.1

This gene is predicted to code for threonine synthase. A NCBI BLAST search gave threonine synthases as top hits [[2]](http://198.128.28.118/genome/converter.php#ref-2) with e-values of 1E-92 to 1E-31 and around 28% to 39% identity. A PDB BLAST search gave as a top hit an article on threonien synthase crystal structure [[17]](http://198.128.28.118/genome/converter.php#ref-17) with an e-value of 2.49267E-47 and 36% identity. Out of the active sites mentioned in the article, 9 out of 19 of them matched. The mismatches were T88L, S155N, V156L, R160Y, V186L, G187A, N188S, A189G, G190S and N191L. Some of these mismatches are not major due to amino acid similarity, namely the V156L, V186L, G187A and A189G mismatches. T88 was mentioned as interacting with an ester oxygen via its hydroxyl group, while L88 has no such group. S155 was said to interact with a phosphate via its side chain. S155 has a polar hydroxyl group, while N155 has a polar carboxamide group. R160 was said to interact with a phosphate via its side chain as well. R160 has many nitrogens to accomplish this, while Y160 has a phenol group. However, R160 and Y160 are quite different in structure. N188 interacts with a phosphate with its side chain too; N188 has a polar caboxamide group, while S188 has a polar hydroxyl group. G190 and S190 are dissimilar because one is nonpolar and hydrophobic while the other is polar and hydrophilic. N191 helps to bind a phosphate group; it has a polar carboxamide group, but L191 has no such polar group. A reverse BLAST hit gave not this gene but the gene annotated below as the top hit. From all these data, the gene is still predicted to code for threonine synthase, but more studies need to be done to confirm this.

Gene OID: 2505166218 EC: 4.2.3.1

This gene is predicted to code for threonine synthase. A NCBI BLAST search gave as top hits other threonine synthases [[13]](http://198.128.28.118/genome/converter.php#ref-13) with e-values ranging from 0 to 8E-97 and

50% to 75% identity. A PDB BLAST search gave as a top hit an article on threonine synthase crystal structure [[17]](http://198.128.28.118/genome/converter.php#ref-17) with an e-value of 6.3999E-113 and 60% identity. Out of the active sites mentioned, 19 out of 19 of them matched, and a reverse BLAST search gave this gene as the top hit. Therefore, this gene is strongly predicted to code for threonine synthase.

**Homoserine O-succinyltransferase**

EC: 2.3.1.46

No gene was predicted by IMG/ER to code for this protein. A search was done on UniProt for cyanobacterial FASTA sequences for this protein, and a BLAST search was done against *Microcoleus vaginatus* genome in IMG/ER. This BLAST search gave no good hits as results. There were hits, but they had e-values of 2 or greater and were not labeled as homoserine O-succinyltransferases. It is predicted then that *Microcoleus vaginatus* does not express this protein.

**O-succinylhomoserine lyase**

EC: 2.5.1.48

No gene was predicted by IMG/ER to code for this protein. A search was done on UniProt for cyanobacterial FASTA sequences for this protein, and a BLAST search was done against *Microcoleus vaginatus* genome in IMG/ER. This BLAST search gave no good hits as results. There were hits, but they had e-values of 2 or greater and were not labeled as O-succinylhomoserine lyase. It is predicted then that *Microcoleus vaginatus* does not express this protein.

**Cystathionine beta-lyase**

EC: 4.4.1.8

No gene was predicted by IMG/ER to code for this protein. A search was done on UniProt for cyanobacterial FASTA sequences for this protein, but no such FASTA sequences were found. Instead, bacterial FASTA sequences were used for comparison and a BLAST search was done against *Microcoleus vaginatus* genome in IMG/ER. This BLAST search gave no good hits as results. There were hits, but they had e-values of 2 or greater and were not labeled as cystathionine beta-lyase. It is predicted then that *Microcoleus vaginatus* does not express this protein.

**Methionine synthase**

Gene OID: 2505170425 EC: 2.1.1.13

This gene is predicted to code for methionine synthase. A NCBI BLAST search gave other methionine synthases as top hits [[5]](http://198.128.28.118/genome/converter.php#ref-5) with e-values from 0 to 1E-94 and from around 34% to 76% identity. A PDB BLAST search gave as a top hit an article on the crystal structure of methionine synthase [[1]](http://198.128.28.118/genome/converter.php#ref-1) with an e-value of 1.48768E-58 and 29% identity. 7 out of 13 of the active sites mentioned in this article matched up. The mismatches were P1137F, H1145M, A1170Q, M1171L, G1174E and S1178T. The paper did not specifically mentioned what the function of each of these groups was. In light of this, a multiple sequence alignment was done on T-Coffee using cyanobacterial FASTA sequences from UniProt. This showed that at site 1137, the F1137 residue was completely conserved among all cyanobacteria. At site 1145, some other cyanobacteria had the M1145 residue, but none had the H1145 residue. The Q1170 residue was completely conserved among cyanobacteria. Except for one cyanobacteria with a V1171 residue, the L1171 was also completely conserved among cyanobacteria. The E1174 residue as well was completely conserved, and so was the T1178 residue. This showed that the mismatches were justified since they are observed in other cyanobacteria as well. Additionally, a reverse BLAST search gave this gene as the first hit. This gene is predicted to code for methionine synthase.

**Cobalamin-independent homocysteine transmethylase**

EC: 2.1.1.14

No gene was predicted by IMG/ER to code for this protein. A search was done on UniProt for cyanobacterial FASTA sequences for this protein, and a BLAST search was done against *Microcoleus vaginatus* genome in IMG/ER. This BLAST search gave no good hits as results. There were hits, but they had e-values of 2 or greater and were not labeled as cobalamin-independent homocysteine transmethylase. It is predicted then that *Microcoleus vaginatus* does not express this protein.

**LL-diaminopimelate aminotransferase**

Gene OID: 2505166565 EC: 2.6.1.83

This gene is predicted to code for L,L-diaminopimelate aminotransferase. A NCBI BLAST search gave as top hits other L,L-diaminopimelate aminotransferases {{NCBI: Q8YM38}}. The top 10 hits were especially good hits, with all of them having an e-value of 0 and around 85% identity. A PDB BLAST search gave an article on L,L-diaminopimelate aminotransferase crystal structure [[8]](http://198.128.28.118/genome/converter.php#ref-8) with an e-value of 1.50525E-110 and 52% identity. Out of the active sites mentioned in the article, 12 of 13 matched, with the lone mismatch being Y283N. Y283 was not mentioned in the article as having specific catalytic purpose, so this mismatch does not seem to be major. Additionally, a reverse BLAST search gave this gene as the top hit, so the gene is predicted to code for L,L-diaminopimelate aminotransferase.

**Homoserine O-acetyltransferase**

EC: 2.3.1.31

No gene was predicted by IMG/ER to code for this protein. A search was done on UniProt for cyanobacterial FASTA sequences for this protein, and a BLAST search was done against *Microcoleus vaginatus* genome in IMG/ER. This BLAST search gave no good hits as results. There were hits, but they had e-values of 2 or greater and were not labeled as homoserine O-acetyltransferase. It is predicted then that *Microcoleus vaginatus* does not express this protein.

**O-acetyl-L-homoserine sulfhydrylase**

EC: 2.5.1.49

No gene was predicted by IMG/ER to code for this protein. A search was done on UniProt for cyanobacterial FASTA sequences for this protein, and a BLAST search was done against *Microcoleus vaginatus* genome in IMG/ER. This BLAST search gave no good hits as results. There were hits, but they had e-values of 2 or greater and were not labeled as O-acetyl-L-homoserine sulfhydrylase. It is predicted then that *Microcoleus vaginatus* does not express this protein.

**O-succinylhomoserine sulfhydrylase**

EC: 2.5.1.-

No gene was predicted by IMG/ER to code for this protein. A search was done on UniProt for cyanobacterial FASTA sequences for this protein, and a BLAST search was done against *Microcoleus vaginatus* genome in IMG/ER. This BLAST search gave no good hits as results. There were hits, but they had e-values of 2 or greater and were not labeled as O-succinylhomoserine sulfhydrylase. It is predicted then that *Microcoleus vaginatus* does not express this protein.

**Functional Studies**

Some genes annotated in this pathway were not definitively coding for proteins, while other genes coded for the same protein as another gene. In both of these cases, additionally functional studies should be done on those genes, in order to better understand what the genes do code for and what they do not code for.

Functional studies should be done on gene 2505168419. This gene is predicted to code for diaminopimelate decarboxylyase. However, gene 2505169231 also is predicted to code for diaminopimelate decarboxylase, and with greater certainty. Therefore, functional studies should be done to see if gene 2505168419 codes for the protein, first of all. If it does, additional studies should be done to see why two genes code for proteins with the same function, and if there is an evolutionary benefit to having two genes coding for this one protein instead of only having one gene coding for the protein. It should also be seen if the different genes are expressed under certain different external circumstances.

Similar to the gene mentioned previously, gene 2505165746 should have functional studies done on it. This gene is predicted to code for threonine synthase, but not with strong certainty. This needs to be checked. If indeed the gene codes for threonine synthase, additional studies should be done to see why this gene and gene 2505166218 both code for proteins with the same function, similar to the pair of genes mentioned in the previous paragraph.

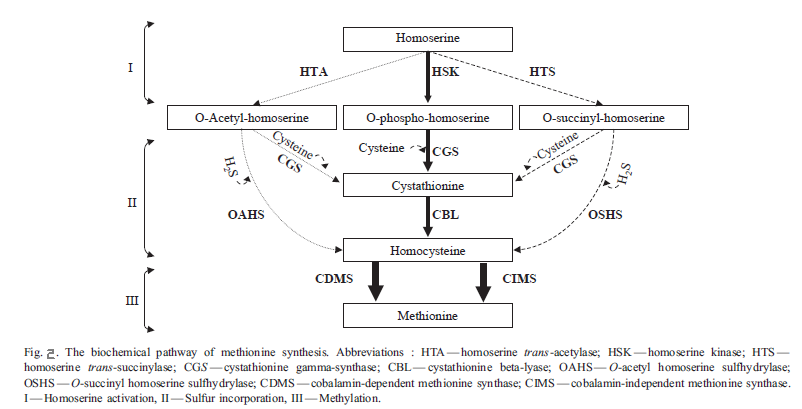
Gene 2505169061 (predicted to code for homoserine kinase) also should have functional studies done on it. For this gene, there was bad active site matching, casting doubt on whether or not this gene actually codes for homoserine kinase or not. An experiment should be done with this gene knocked out, and it should be seen if homoserine kinase is still present in the organism.

**Discussion**

*Microcoleus vaginatus* was found to be missing many crucial enzymes for methionine and lysine biosynthesis. Referring to S Figure 1, *Microcoleus vaginatus* was found to be missing tetrahydrodipicolinate succinylase (EC: 2.3.1.117) and N-succinyl-L-diaminopimelate desuccinylase (EC: 3.5.1.18) along the lysine pathway, and it was missing homoserine O-succinyltransferase (EC: 2.3.1.46), O-succinlhomoserine-lyase (EC: 2.5.1.48) and cystathionine-beta-lyase (EC: 4.4.1.8). Due to the absence of these enzymes, *Microcoleus vaginatus* is unable to synthesize its own methionine or lysine, at least according to this pathway. It might be the case that *Microcoleus vaginatus* uses a different pathway. A literature search was done, and an article [[11]](http://198.128.28.118/genome/converter.php#ref-11) mentioned an alternate pathway that cyaonbacteria could use for lysine biosynthesis.

This alternate pathway involves LL-diaminopimelate aminotransferase (EC: 2.6.1.83), which converts (S)-2,3,4,5-tetrahydrodipicolinate along with L-glutamate straight into L,L-diaminopimelate and 2-oxoglutarate, skipping over the steps that require enzyme 2.3.1.117 and enzyme 3.5.1.18. *Microcoleus vaginatus* then has the enzymes to ultimately convert L,L-diaminopimelate into L-lysine. In this way *Microcoleus vaginatus* can still synthesize its own lysine.

For methionine biosynthesis, there were two alternate pathways that were found {{Pubmed: 16046084}}, but the enzymes mentioned for these alternate pathways were not found. An illustration of the alternate pathways is shown below.



Referring to S Figure 2, the pathway shown on the MetaCyc map is the right path that uses HTS, CGS, and CBL. None of these enzymes were predicted to exist though, so *Microcoleus vaginatus* cannot use this pathway to synthesize methionine. The middle path looked promising, because *Microcoleus vaginatus* has homoserine kinase, but again it does not have CGS or CBL and so cannot get to the biosynthesis of methionine. Finally, along the left path, HTA, CGS and OAHS were not found in *Microcoleus vaginatus*, so again this does not allow for the synthesis for methionine. It is interesting to note that CDMS, the enzyme catalyzing the reaction of homocysteine to methionine, does exist in *Microcoleus vaginatus*. It is interesting that an organism would have this enzyme without any way to synthesize homocysteine.

This result is quite troubling, since *Microcoleus vaginatus* is known to be able to grow without any amino acid supplements, which means that the bug must synthesize all of its own amino acids. *Microcoleus vaginatus* must have some alternate way of synthesizing methionine, or it has an alternate pathway for synthesizing homocysteine. The latter option seems more viable, because the bug already has the enzyme to convert homocysteine to methionine; it just needs a way to synthesize homocysteine.

In conclusion, *Microcoleus vaginatus* is predicted to be able to synthesize lysine and threonine but not methionine. It is able to synthesize lysine via an alternate pathway, but is unable to synthesize methionine because it is missing enzymes. However, it is known that *Microcoleus vaginatus* can grow without amino acid sources, which suggests that the it can synthesize methionine. Additional research should be done to see what this pathway is that methionine is synthesized by, because there is a good chance that it could be a new pathway.

**References**

[1] Bandarian V, Pattridge KA, Lennon BW, Huddler DP, Matthews RG, Ludwig ML. "Domain

alternation switches B(12)-dependent methionine synthase to the activation

conformation." Nat Struct Biol. 2002 Jan; 9(1):53-6.

[2] Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM,

Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI,

Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen

NS, Venter JC. "Complete genome sequence of the methanogenic archaeon,

Methanococcus jannaschii." Science. Aug 23, 1996; 273(5278):1058-73.

[3] Cirilli M, Zheng R, Scapin G, Blanchard JS. "The three-dimensional structures of the

Mycobacterium tuberculosis dihydrodipicolinate reductase-NADH-2,6-PDC and

-NADPH-2,6-PDC complexes. Structural and mutagenic analysis of relaxed nucleotide

specificity." Biochemistry. Sep 16, 2003; 42(36):10644-50.

[4] Clepet C, Borne F, Krishnapillai V, Baird C, Patte JC, Cami B. "Isolation, organization and

expression of the Pseudomonas aeruginosa threonine genes." Mol Microbiol. 1992 Nov;

6(21):3109-19.

[5] Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas

S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R,

Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K,

Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream

MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor

K, Whitehead S, Barrell BG. "Deciphering the biology of Mycobacterium tuberculosis

from the complete genome sequence." Nature. Jun 11, 1998; 393(6685):537-44.

[6] DeLaBarre B, Thompson PR, Wright GD, Berghuis AM. "Crystal structures of homoserine

dehydrogenase suggest a novel catalytic mechanism for oxidoreductases." Nat Struct

Biol. 2000 Mar; 7(3):238-44.

[7] Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R,

Snead MA, Keller M, Aujay M, Huber R, Feldman RA, Short JM, Olsen GJ, Swanson

RV. "The complete genome of the hyperthermophilic bacterium Aquifex aeolicus."

Nature. Mar 26, 1998; 392(6674):353-8.

[8] Dobson RC, Girón I, Hudson AO. "L,L-diaminopimelate aminotransferase from

Chlamydomonas reinhardtii: a target for algaecide development." PLoS One. 2011;

6(5):e20439.

[9] Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, Barbe V, Duprat S,

Galperin MY, Koonin EV, Le Gall F, Makarova KS, Ostrowski M, Oztas S, Robert C,

Rogozin IB, Scanlan DJ, Tandeau de Marsac N, Weissenbach J, Wincker P, Wolf YI,

Hess WR. "Genome sequence of the cyanobacterium Prochlorococcus marinus SS120,

a nearly minimal oxyphototrophic genome." Proc Natl Acad Sci U S A. Aug 19, 2003;

100(17):10020-5.

[10] Gokulan K, Rupp B, Pavelka MS Jr, Jacobs WR Jr, Sacchettini JC. "Crystal structure of

Mycobacterium tuberculosis diaminopimelate decarboxylase, an essential enzyme in

bacterial lysine biosynthesis." J Biol Chem. May 16, 2003; 278(20):18588-96.

[11] Gophna U, Bapteste E, Doolittle WF, Biran D, Ron EZ. "Evolutionary plasticity of methionine

biosynthesis." Gene. Aug 1, 2005; 355():48-57.

[12] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[13] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N,

Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda

M, Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[14] Kaneko T, Tanaka A, Sato S, Kotani H, Sazuka T, Miyajima N, Sugiura M, Tabata S.

"Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp.

strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to

92% of the genome." DNA Res. Aug 31, 1995; 2(4):153-66, 191-8.

[15] Lushy A, He Z, Fish A, Darash-Yahana M, Minai L, Verchovsky L, Michaeli D, Nechushtai

R. "An insight into the assembly and organization of photosystem I complex in the

thylakoid membranes of the thermophilic cyanobacterium, Mastigocladus laminosus."

Indian J Biochem Biophys. 2000 Dec; 37(6):405-17.

[16] Nakai T, Okada K, Akutsu S, Miyahara I, Kawaguchi S, Kato R, Kuramitsu S, Hirotsu K.

"Structure of Thermus thermophilus HB8 aspartate aminotransferase and its complex

with maleate." Biochemistry. Feb 23, 1999; 38(8):2413-24.

[17] Omi R, Goto M, Miyahara I, Mizuguchi H, Hayashi H, Kagamiyama H, Hirotsu K. "Crystal

structures of threonine synthase from Thermus thermophilus HB8: conformational

change, substrate recognition, and mechanism." J Biol Chem. Nov 14, 2003;

278(46):46035-45.

[18] Parsot C, Mazel D. "Cloning and nucleotide sequence of the thrB gene from the

cyanobacterium Calothrix PCC 7601." Mol Microbiol. 1987 Jul; 1(1):45-52.

[19] Pillai B, Moorthie VA, van Belkum MJ, Marcus SL, Cherney MM, Diaper CM, Vederas JC,

James MN. "Crystal structure of diaminopimelate epimerase from Arabidopsis thaliana,

an amino acid racemase critical for L-lysine biosynthesis." J Mol Biol. Jan 16, 2009;

385(2):580-94.

[20] Ray SS, Bonanno JB, Rajashankar KR, Pinho MG, He G, De Lencastre H, Tomasz A,

Burley SK. "Cocrystal structures of diaminopimelate decarboxylase: mechanism,

evolution, and inhibition of an antibiotic resistance accessory factor." Structure. 2002

Nov; 10(11):1499-508.

[21] Robin AY, Cobessi D, Curien G, Robert-Genthon M, Ferrer JL, Dumas R. "A new mode of

dimerization of allosteric enzymes with ACT domains revealed by the crystal structure of

the aspartate kinase from Cyanobacteria." J Mol Biol. Jun 4, 2010; 399(2):283-93.

[22] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[23] Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh

R, Blakely D, Cook R, Gilbert K, Harrison D, Hoang L, Keagle P, Lumm W, Pothier B, Qiu

D, Spadafora R, Vicaire R, Wang Y, Wierzbowski J, Gibson R, Jiwani N, Caruso A, Bush

D, Reeve JN. "Complete genome sequence of Methanobacterium thermoautotrophicum

deltaH: functional analysis and comparative genomics." J Bacteriol. 1997 Nov;

179(22):7135-55.

[24] Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FS,

Hufnagle WO, Kowalik DJ, Lagrou M, Garber RL, Goltry L, Tolentino E,

Westbrock-Wadman S, Yuan Y, Brody LL, Coulter SN, Folger KR, Kas A, Larbig K, Lim

R, Smith K, Spencer D, Wong GK, Wu Z, Paulsen IT, Reizer J, Saier MH, Hancock RE,

Lory S, Olson MV. "Complete genome sequence of Pseudomonas aeruginosa PAO1, an

opportunistic pathogen." Nature. Aug 31, 2000; 406(6799):959-64.

[25] Ura H, Nakai T, Kawaguchi SI, Miyahara I, Hirotsu K, Kuramitsu S. "Substrate recognition

mechanism of thermophilic dual-substrate enzyme." J Biochem. 2001 Jul; 130(1):89-98.

[26] Viola RE, Liu X, Ohren JF, Faehnle CR. "The structure of a redundant enzyme: a second

isoform of aspartate beta-semialdehyde dehydrogenase in Vibrio cholerae." Acta

Crystallogr D Biol Crystallogr. 2008 Mar; 64(Pt 3):321-30.

[27] Voss JE, Scally SW, Taylor NL, Atkinson SC, Griffin MD, Hutton CA, Parker MW, Alderton

MR, Gerrard JA, Dobson RC, Dogovski C, Perugini MA. "Substrate-mediated

stabilization of a tetrameric drug target reveals Achilles heel in anthrax." J Biol Chem.

Feb 19, 2010; 285(8):5188-95.

[28] Weyand S, Kefala G, Weiss MS. "The three-dimensional structure of

N-succinyldiaminopimelate aminotransferase from Mycobacterium tuberculosis." J Mol

Biol. Mar 30, 2007; 367(3):825-38.

[29] Yamamoto J, Shimizu M, Yamane K. "Nucleotide sequence of the

diaminopimelate-decarboxylase gene from Bacillus subtilis." Nucleic Acids Res. Dec 11,

1989; 17(23):10105.

[30] Zhou T, Daugherty M, Grishin NV, Osterman AL, Zhang H. "Structure and mechanism of

homoserine kinase: prototype for the GHMP kinase superfamily." Structure. Dec 15,

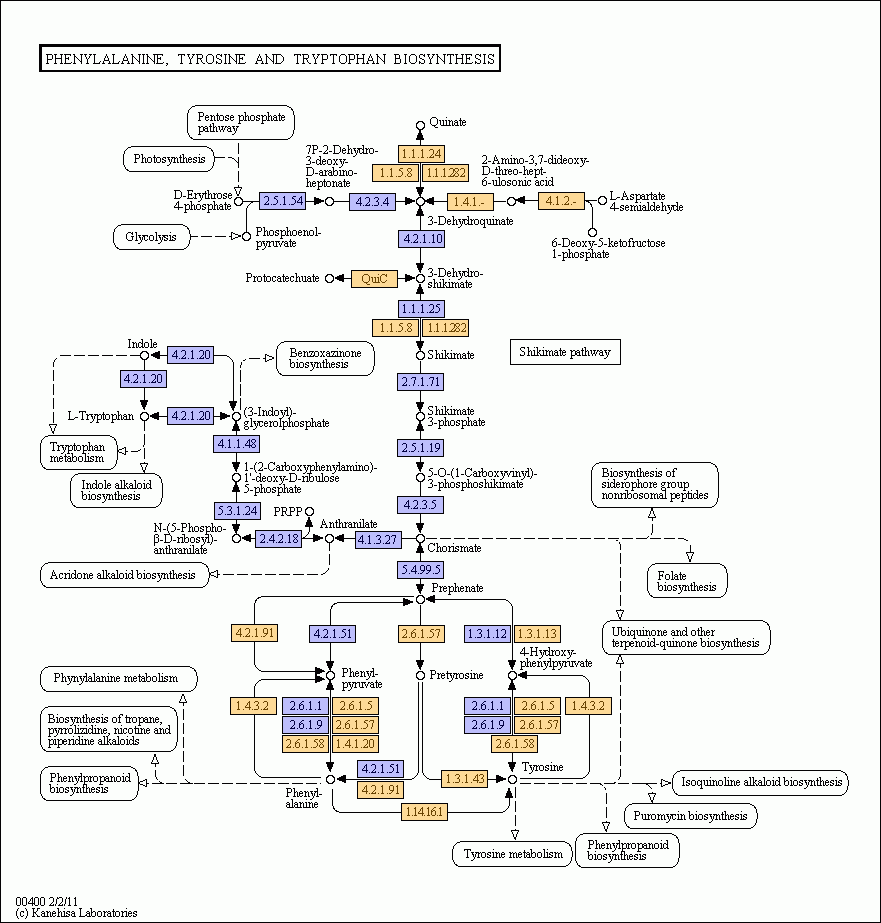
2000; 8(12):1247-57.

## Phenylalanine, Tyrosine and Tryptophan

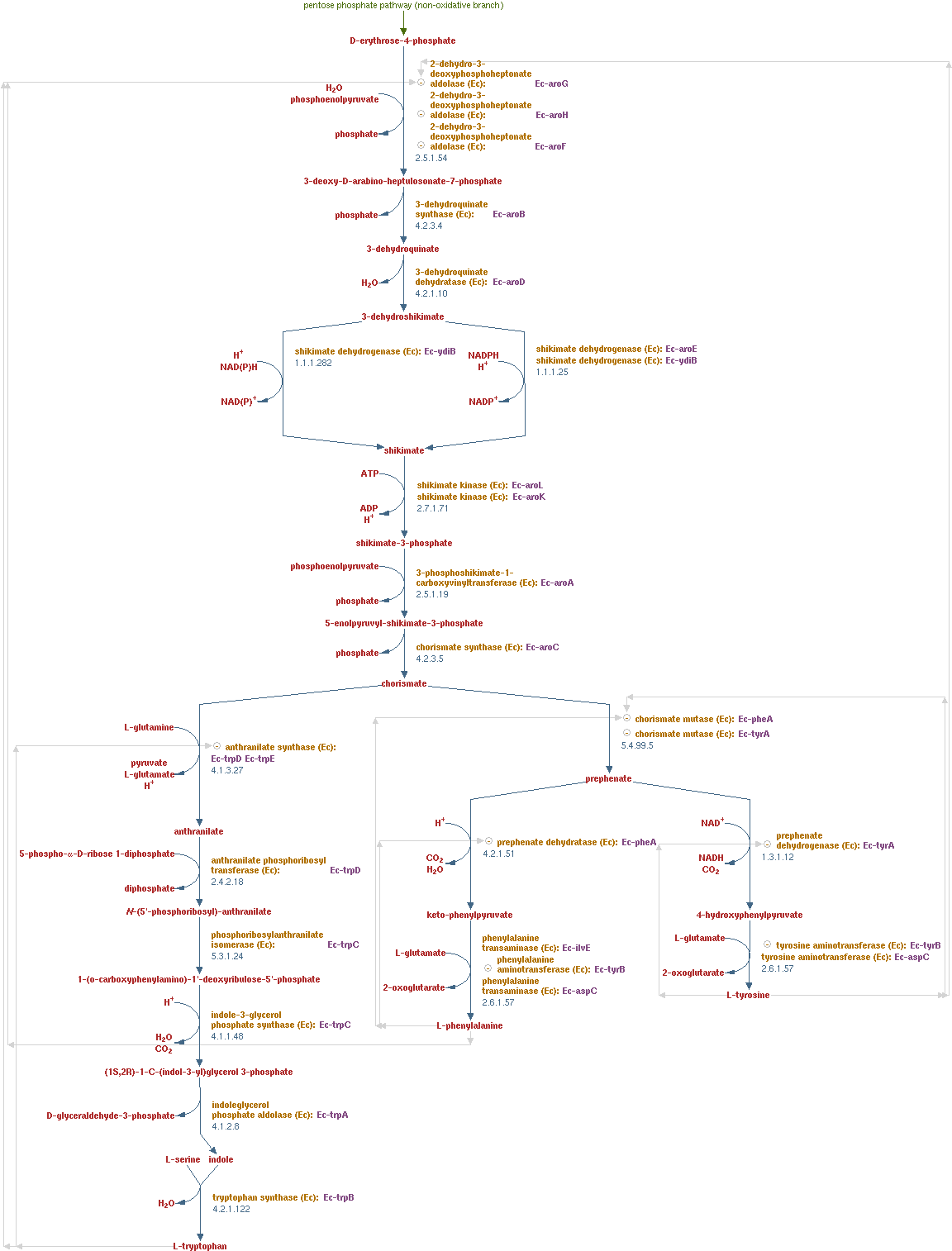
Phenylalanine (F), tyrosine (Y) and tryptophan (W) are the three aromatic amino acids. They share a common superpathway since they are all aromatic, much like the branched amino acids leucine, isoleucine and valine. F and W are essential amino acids for humans, while Y is non-essential. It seems reasonable that *Microcoleus vaginatus* would produce F and W on its own unlike humans, because *Microcoleus vaginatus* cannot obtain these amino acids by ingesting other organisms like humans can.

The aromatic amino acids L-phenylalanine, L-tyrosine and L-tryptophan are biosynthesized from the principal common precursor chorismate. The conversion of chorismate to prephenate leads to the terminal pathways of L-phenylalanine and L-tyrosine biosynthesis. The conversion of chorismate to anthranilate leads to the biosynthesis of L-tryptophan. *Microcoleus vaginatus* is predicted to be able to undergo biosynthesis of all 3 of these amino acids.

**Automated Annotation**



S Fig. 1: KEGG pathway map of F, Y, W Amino Acid Biosynthesis. Blue boxes indicate genes predicted by IMG/ER and orange boxes indicate genes not predicted to be in *MIcrocoleus vaginatus* by IMG/ER.



S Fig. 2: MetaCyc pathway map of F, Y, W Amino Acid Biosynthesis. Full-scale image at <http://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=COMPLETE-ARO-PWY&detail-level=2&detail-level=1>

**Phospho-2-dehydro-3-deoxyheptonate aldolase**

Two genes were predicted by IMG/ER to code for this protein. Both genes were predicted to have function for the protein.

Gene OID: 2505166457 EC: 2.5.1.54

This gene was predicted to code for phospho-2-dehydro-3-deoxyheptonate aldolase. A NCBI BLAST search gave as the top two hits phospho-2-dehydro-3-deoxyheptonate aldolases [[30]](http://198.128.25.103/genome/converter.php#ref-30) with e-values of 1E-96 and 5E-72, and with 47% and 51% identity. A PDB BLAST search gave an article on the crystal structure of the protein [[26]](http://198.128.25.103/genome/converter.php#ref-26) with an e-value of 3.38372E-76 and 47% identity. 15 out 15 of the active sites mentioned in this article matched up, and a reverse BLAST search gave gene 2505166457 as the top hit, so this gene is strongly predicted to code for the protein.

Gene OID: 2505170171 EC: 2.5.1.54

This gene was also predicted to code for phospho-2-dehydro-3-deoxyheptonate aldolase. A NCBI BLAST search gave as the two top hits phospho-2-dehydro-3-deoxyheptonate aldolases [[30]](http://198.128.25.103/genome/converter.php#ref-30) with e-values of 3E-90 and 3E-87, and with 49% identity. A PDB BLAST search gave the same article on the crystal structure of phospho-2-dehydro-3-deoxyheptonate alodlase [[26]](http://198.128.25.103/genome/converter.php#ref-26), with an e-value of 5.48422E-71 and 48% identity. Out of the active sites mentioned, 14 out of 15 of them matched. The sole mismatch was A185S. This is not a major mismatch, because the article mentioned it only as being important for its main chain N atom, which all amino acids have. Also, the PDB sequence aligner considers A and S as similar residues. A reverse BLAST search gave gene 2505166457 as the top hit but this gene as the next hit. From this data, it is concluded that this gene codes for phospho-2-dehydro-3-deoxyheptonate aldolase, but with a little less certainty than the previous gene.

**3-dehydroquinate synthase**

Gene OID: 2505168245 EC: 4.2.3.4

This gene is predicted to code for 3-dehydroquinate synthase. A NCBI BLAST search gave only 3-dehydroquinate synthases as hits [[10]](http://198.128.25.103/genome/converter.php#ref-10), with the best hit having an e-value of 0 and the worst hit having an e-value of 2E-96, and percent identities from 43% to 71%. A PDB BLAST search gave an article on 3-dehydroquinate synthase as the top hit [[17]](http://198.128.25.103/genome/converter.php#ref-17) with an e-value of 3.66416E-65 and 41% identity. Out of the active sites mentioned in the article, 12 out of 12 of them matched, and a reverse BLAST search gave this gene as the first hit. Therefore, this gene is strongly predicted to code for 3-dehydroquinate synthase.

**3-dehydroquinate dehydratase**

Gene OID: 2505168411 EC: 4.2.1.10

This gene is predicted to code for 3-dehydroquinate dehydratase. A NCBI BLAST search gave only 3-dehydroquinate dehydratases as hits {{NCBI: Q8YTE0}} with e-values ranging from 4E-77 to 5E-44 and 48% to 78% identity. A PDB BLAST search gave an article on the crystal structure of 3-dehydroquinate dehydratase [[18]](http://198.128.25.103/genome/converter.php#ref-18) with an e-value of 4.57422E-35 and 51% identity. Out of the active sites mentioned, 8 out of 9 of them matched, with the mismatch being E55G. The article mentioned that in some different 3-dehydroquinate dehydratases, the E55 can be replaced by G55. Additionally, a reverse BLAST search gave the same gene as the top hit. From these data, the gene is predicted to code for 3-dehydroquinate dehydratase.

**Shikimate 5-dehydrogenase**

Gene OID: 2505169604 EC: 1.1.1.25

This gene is predicted to code for shikimate 5-dehydrogenase. A NCBI BLAST search gave only shikimate 5-dehydrogenases as hits {{NCBI: Q8YVC1}}, with e-values from 4E-139 to 3E-36, and 29% to 62% identity. A PDB BLAST search gave an article on shikimate 5-dehydrogenase [[23]](http://198.128.25.103/genome/converter.php#ref-23) with an e-value of 2.43197E-41 and 35% identity. Out of the active sites mentioned, 12 out of 14 of them matched. The mismatches were A131N and K157L. At site 131, A131 was only important for its main chain nitrogen donating a hydrogen bond. Any other amino acid could take this role, like N131. At site 157, the article mentioned hydrogen bonding being important. K157 can hydrogen bond with its side group, since there is an amide group. L157 cannot hydrogen bond with its (nonpolar) side group, but it can still hydrogen bond with its main group nitrogen. Also, a reverse BLAST search gave this gene as the top hit. From all these data, it is predicted that the gene codes for shikimate 5-dehydrogenase.

**Shikimate kinase**

Gene OID: 2505168597 EC: 2.7.1.71

This gene is predicted to code for shikimate kinase. A NCBI BLAST search gave only shikimate kinases as hits {{NCBI: Q8YXG9}} with e-values from 2E-66 to 6E-27 and 37% to 53% identity. A PDB BLAST search gave an article on shikimate kinase as a top hit [[4]](http://198.128.25.103/genome/converter.php#ref-4),

with an e-value of 1.09854E-32 and 42%. Out of the active sites mentioned in the article, 13 out of 16 of them matched. The mismatches were S267G, S269G and D270E. The article did not mention these residues as being especially special or essential for enzyme function. A multiple sequence alignment was done on T-Coffee to double check, comparing the query and subject genes to other reviewed cyanobacterial genes from UniProt. For sites 267 and 269, there was no alignment with the other cyanobacteria that could be forced here. It seems though that the mismatches are reasonable. Both the S residues are changed to G residues, and the only difference between the residues is a hydroxyl group. For site 270, the alignment showed that other cyanobacteria had E270 instead of D270, and the D270 and E270 residues are very similar. They only differ by an extra carbon chain in their side chains. Additionally, a reverse BLAST search gave the same gene as the top hit, so it is predicted that the gene codes for shikimate kinase.

**3-phosphoshikimate 1-carboxyvinyltransferase**

Gene OID: 2505170576 EC: 2.5.1.19

This gene is predicted to code for 3-phosphoshikimate 1-carboxyvinyltransferase. A NCBI BLAST search gave only 3-phosphoshikimate 1-carboxyvinyltransferases as hits [[25]](http://198.128.25.103/genome/converter.php#ref-25) with e-values from 0 to 1E-125 and 49% to 76% identity. A PDB BLAST search gave an article on 3-phosphoshikimate 1-caarboxyvinyltransferase [[24]](http://198.128.25.103/genome/converter.php#ref-24) as a top hit, with an e-value of 3.17221E-101 and 49% identity. Out of the active sites mentioned, 16 out of 16 of them matched, and a reverse BLAST search gave the gene as the top hit. Therefore, the gene is strongly predicted to code for 3-phosphoshikimate 1-carboxyvinyltransferase.

**Chorismate synthase**

Gene OID: 2505168294 EC: 4.2.3.5

This gene is predicted to code for chorismate synthase. A NCBI BLAST search gave only chorismate synthases as top hits {{NCBI: Q8YYP9}} with e-values from 0 to 2E-120, and 50% to 87% identity. A PDB BLAST search gave as a top hit an article on chorismate synthase [[1]](http://198.128.25.103/genome/converter.php#ref-1) with an e-value of 2.02131E-71 and 44% identity. Out of the active sites mentioned, 11 out of 15 of them matched. The mismatches were N241P, P299A, S300T and I327V. Site 241 was mentioned as important since the amine group on its side chain can hydrogen bond. P241 has a N that it can also use to hydrogen bond. Site 299, P299 is important for its hydrophobic property. Although A299 is not similar to P299 in structure, it is also hydrophobic. S300 and I327 are also important for their hydrophobic properties. T300 and V327 are similar in that they have similar degrees of hydrophobicity to S300 and I327. Additionally, a reverse BLAST search gave this gene as the top hit. It is therefore predicted that the gene codes for chorismate synthase.

**Anthranilate synthase**

This protein has two genes predicted to have function for it by IMG/ER. This protein is a complex that has two subunits. It is hypothesized that the two genes annotated below correspond to the two subunits. Either way, both of the subunits are predicted to exist, so the complex as a whole is expected to exist.

Gene OID: 2505171067 EC: 4.1.3.27

This gene is predicted to code for anthranilate synthase. A NCBI BLAST search gave other anthranilate synthases (component I) as top hits [[21]](http://198.128.25.103/genome/converter.php#ref-21), with e-values from 0 to 3E-54 and 44% to 72% identity. A PDB BLAST search gave an article on anthranilate synthase as a top hit[[20]](http://198.128.25.103/genome/converter.php#ref-20) with an e-value of 5.82139E-54 and 31% identity. Out of the active sites mentioned, 8 out of 11 of them matched. The mismatches were K50Y, F294A and C465A. The article showed that K50 donated a hydrogen bond with the amine group on its side chain. Y50 could also donate a hydrogen bond, using the hydroxyl group on its side chain. Site 294 is not mentioned to have special function, but both F294 and A294 are nonpolar and hydrophobic, so both could contribute hydrophobic interactions. C465 is mentioned for its hydrophobic interaction, which A465 can contribute too. Also, a reverse BLAST search gave this gene as the top hit, so the gene is predicted to code for anthranilate synthase.

Gene OID: 2505171087 EC: 4.1.3.27

This gene is predicted to code for anthranilate synthase as well. A NCBI BLAST search gave other anthranilate synthases as top hits [[13]](http://198.128.25.103/genome/converter.php#ref-13) with e-values from 3E-85 to 1E-22 and around 50% to 60% identity. A PDB BLAST search gave an article on anthranilate synthase as a top hit [[14]](http://198.128.25.103/genome/converter.php#ref-14), with an e-value of 3.17153E-44 and 48% identity. 9 active sites were mentioned in the article, but only 3 of them could be checked because this gene had a FASTA sequence shorter than that of the subject. Out of those 3 active sites, all of them matched. A reverse BLAST search also gave this gene as the top hit. This gene is predicted to code for component II since the other gene was predicted to code for component I.

**Anthranilate phosphoribosyltransferase**

There were two genes predicted by IMG/ER to code for this protein. Gene 2505168075 could possibly be an anthranilate phosphoribosyltransferase, while gene 2505168983 is strongly predicted to be an anthranilate phosphoribosyltransferase.

Gene OID: 2505168075 EC: 2.4.2.18

This gene may or may not code for anthranilate phosphoribosyltransferase. A NCBI BLAST search gave only anthranilate phosphoribosyltransferases as hits but for one hit {{NCBI: A1KAT3}}. The e-values of the hits ranged from 3E-26 to 9E-14, with 24% to 30% identity. A PDB BLAST search gave an article on anthranilate phosphoribosyltransferase as a top hit [[15]](http://198.128.25.103/genome/converter.php#ref-15), with an e-value of 7.71003E-12 and 30% identity. Out of the active sites mentioned in the article, only 3 out of 15 matched. This makes it seem that the gene might have been misannotated. Further research needs to be done, because the NCBI BLAST hits suggest that the gene codes for anthranilate phosphoribosyltransferase, but the lack of active site matching contradicts this.

Gene OID: 2505168983 EC: 2.4.2.18

This gene is predicted to code for anthranilate phosphoribosyltransferase. A NCBI BLAST search gave only anthranilate phosphoribosyltransferases as hits {{NCBI: Q8YXQ9}} with e-values from 1E-170 to 3E-67 and 37% to 69% identity. A PDB BLAST search gave an article on anthranilate phosphoribosyltransferase structure [[12]](http://198.128.25.103/genome/converter.php#ref-12) with an e-value of 6.23052E-52 and 36% identity. Out of the active sites mentioned, 9 out of 10 of them matched. The sole mismatch was S99T. S99 was not specifically mentioned in the article, and S99 and T99 are very similar in structure. Additionally, a reverse BLAST search gave the gene as the top hit, so this gene is strongly predicted to code for anthranilate phosphoribosyltransferase.

**Phosphoribosylanthranilate isomerase**

Gene OID: 2505167588 EC: 5.3.1.24

This gene is predicted to code for phosphoribosylanthranilate isomerase. A NCBI BLAST search gave almost all N-(5'-phosphoribosyl)-anthranilate isomerases as hits {{NCBI: Q8YLL0}} with e-values from 9E-94 to 3E-28 and 32% to 63% identity. A PDB BLAST search gave an article on phosphoribosylanthranilate isomerase as a top hit [[8]](http://198.128.25.103/genome/converter.php#ref-8) with an e-value of 3.13495E-35 and 43% identity. Out of the active sites mentioned in the article, 4 out of 5 of them matched. The mismatch was G105R. G105 is not mentioned in the article to have a specific purpose. G105 and R105 are also not similar in structure. However, the other data seem to override this one active site mismatch; a reverse BLAST search gives this gene as the top hit. All these data together makes it likely that the gene codes for the enzyme.

**Indole-3-glycerol phosphate synthase**

Gene OID: 2505170751 EC: 4.1.1.48

This gene is predicted to code for indole-3-glycerol phosphate synthase. A NCBI BLAST search gave only Indole-3-glycerol phosphate synthases as hits {{NCBI: B1WQE4}} with e-values from 3E-147 to 1E-59 and 44% to 71% identity. A PDB BLAST search gave as a top hit an article on indole-3-glycerol phosphate synthase [[9]](http://198.128.25.103/genome/converter.php#ref-9) with an e-value of 2.75833E-28 and 35% identity. Out of the active sites mentioned in the article, 12 out of 12 of them matched. Furthermore, a reverse BLAST search gave the same gene as the top hit. Therefore, this gene is strongly predicted to code for indole-3-glycerol phosphate synthase.

**Tryptophan synthase**

This protein complex has two subunits, alpha and beta. The complex as a whole catalyzes an overall reaction. The subunits each catalyze reactions that are part of the overall reaction. Together, these reactions make up the overall reaction. Both of the subunits are predicted in *Microcoleus vaginatus*.

The alpha subunit catalyzes the alpha reaction

indole-3-glycerol phosphate ↔ indole + D-glyceraldehyde 3-phosphate

while the beta subunit catalyzes the beta reaction

L-serine + indole ↔ L-tryptophan + H2O

and the overall reaction is

L-serine + indole-3-glycerol phosphate ↔ L-tryptophan + d-glyceraldehyde 3-phosphate + H2O

Alpha Subunit

Gene OID: 2505170965 EC: 4.2.1.20

This gene is predicted to code for the alpha subunit of tryptophan synthase. A NCBI BLAST search gave as hits only tryptophan synthase alpha subunits {{NCBI: B1WNQ1}} with e-values ranging from 4E-146 to 2E-57 and anywhere from 40% to 75% identity. A PDB BLAST search gave as a top hit an article on the alpha subunit of tryptophan synthase [[31]](http://198.128.25.103/genome/converter.php#ref-31) with an e-value of 1.10541E-46 and 42% identity. Two active site residues were mentioned as being important for catalytic activity, and both of these matched. Additionally, a reverse BLAST search gave this gene as the top hit. It is therefore strongly predicted that the gene codes for the alpha subunit of tryptophan synthase.

Beta Subunit

Gene OID: 2505167561 EC: 4.2.1.20

This gene is predicted to code for the beta subunit of tryptophan synthase. A NCBI BLAST search gave only tryptophan synthase beta subunits as hits {{NCBI: Q8YQM6}} with e-values from 0 to 2E-178, and percent identities ranging from 61% to 84%. A PDB BLAST search gave an article on tryptophan synthase beta subunit as a top hit [[16]](http://198.128.25.103/genome/converter.php#ref-16) with an e-value of 3.68186E-133 and 59% identity. Out of the active sites mentioned, 11 out of 17 of them matched up. The mismatches were I16M, E132H, Q270S, F274L, L278M and F281L. Sites 16, 132, 270 and 278 are not mentioned specifically in the article as having specific catalytic function. F274 and F281 are mentioned for their hydrophobicity, but L274 and L281 would be hydrophobic too. As a double-check, a multiple sequence alignment was done on T-Coffee with reviewed cyanobacterial FASTA sequences from UniProt. This showed that at site 16, only a few of the other cyanobacteria had the I16 residue, but the majority had the M16 residue. At site 132, more than half of cyanobacteria had the E132 residue, but somehad the H132 residue as well. At site 270, no other cyanobacteria had the Q270 residue, but some had the S270 residue. At site 274, no other cyanobacteria had the F274 residue, but all had the L274 residue. At site 278, no other cyanobacteria had the L278 residue, but all of them had the M278 residue. At site 281, no other cyanobacteria had F281, but all of them except for one had the L281 residue. Additionally, a reverse BLAST search gave this gene as the top hit. From these data, it is predicted that the gene codes for tryptophan synthase beta subunit.

**Chorismate mutase**

Gene OID: 2505167663 EC: 5.4.99.5

This gene is predicted to code for chorismate mutase. A NCBI BLAST search gave as the top hit a chorismate mutase [[6]](http://198.128.25.103/genome/converter.php#ref-6) with an e-value of 1E-23 and 40% identity. A PDB BLAST search gave as a top hit an article on the crystal structure of chorismate mutase [[2]](http://198.128.25.103/genome/converter.php#ref-2) with an e-value of 1.16518E-18 and 41% identity. Out of the active sites mentioned, only 6 out of 20 matched. Many of these mismatches were amino acids being replaced by similar amino acids. Also, a reverse BLAST search gave the same gene as the top hit. This gene is then possibly predicted to code for chorismate mutase, but further research needs to be done to confirm this.

**Prephenate dehydratase**

Gene OID: 2505170603 EC: 4.2.1.51

This gene is predicted to code for prephenate dehydratase. A NCBI BLAST search gave other prephenate dehydratases as top hits [[7]](http://198.128.25.103/genome/converter.php#ref-7) with e-values from 7E-48 to 3E-27 and around 35% identity. A PDB BLAST search gave an article on prephenate dehydratase as a top hit [[28]](http://198.128.25.103/genome/converter.php#ref-28) with an e-value of 1.69452E-20 and 30% identity. Out of the active sites mentioned, 13 out of 15 of them matched. The two mismatches were D126N and Q234D. D126 was mentioned in that its negative charge caused repulsion, along with E58, that the substrate, prephenate, had to overcome since prephenate is also negatively charged. N126 is very similar to structure to D126 (OH group replaced with NH2 group) and N126 is neutral, so it would cause less repulsion and it could make it easier for the substrate to overcome the negative charged repulsion. The article mentioned that at site 234, in addition to Q234, the D234 residue could be found there too. A reverse BLAST hit gave this gene as the top hit. From all these considerations, the gene is predicted to code for prephenate dehydratase.

**Aspartate/Aromatic aminotransferase**

There are two genes predicted by IMG/ER to code for this multifunctional protein. Reverse BLAST searches then showed 3 more genes predicted to be aromatic aminotransferases, but these genes were not assigned EC numbers by IMG/ER, and only 1 of them bore homology to aromatic aminotransferase. This protein can catalyze many reactions, but we are interested in the reactions of keto-phenylpyruvate and L-glutamate to L-phenylalanine and 2-oxoglutarate, and 4-hydroxyphenylpyruvate + L-glutamate to L-tyrosine and 2-oxoglutarate.

Gene OID: 2505169488 EC: 2.6.1.1

This gene is predicted to code for the aminotransferase. A NCBI BLAST search gave as top hits other aspartate aminotransferases [[11]](http://198.128.25.103/genome/converter.php#ref-11) with e-values from 0 to 4E-31 and around 45% to 65% identity. A PDB BLAST search gave as a top hit an article on aspartate aminotransferase [[19]](http://198.128.25.103/genome/converter.php#ref-19) with an e-value of 1.11E-75 and 41% identity. Out of the active sites mentioned, 10 out of 16 of them matched. The mismatches were A96G, V201I, S232A, F121W, M260G and T264S. At site 96, A96 and G96 are very similar in structure (-CH3 versus -H as side group) and for A96, hydrogen bonding occurs with its [main chain] nitrogen, which can apply for G96 too. At site 201, V201 and I201 are also similar, and V201 was mentioned for its two methyl groups interacting with the substrate; I201 has two methyl groups as well. At site 232, S232 and A232 are not similar, and S232 hydrogen bonds with its side-chain oxygen, while A232 doesn’t have a side chain oxygen. F121 was mentioned as important for its benzene ring in providing hydrophobic interactions; W121 also has a benzene ring (albeit a fused one with N). M260 was mentioned for its hydrophobicity; G260 is hydrophobic as well. T264 and S264 are very similar in structure, and T264 was mentioned for the role it plays in hydrogen bonding with its hydroxyl group on its side chain; S264 also has such a group on its side chain. Finally, a reverse BLAST search gave this gene as the top hit. From these data, it is concluded that the gene codes for an aminotransferase.

Gene OID: 2505171077 EC: 2.6.1.1

This gene is also predicted to code for the aminotransferase. A NCBI BLAST search gave other aminotransferases as top hits [[22]](http://198.128.25.103/genome/converter.php#ref-22) with e-values from 8E-51 to 6E-10 and around 25% to 30% identity. A PDB BLAST search gave as a top hit an article on aspartate aminotransferase [[5]](http://198.128.25.103/genome/converter.php#ref-5) with an e-value of 9.99826E-28 and 29% identity. Out of the active sites mentioned in the article, 5 out of 9 of them matched. The mismatches were Q32I, T88V, F112W and Y194F. Q32 is mentioned as important to form a hydrophobic region, and I32 is hydrophobic as well. The article mentioned that T88 is not necessarily a conserved residue among aminotransferases. F112 is important for its hydrophobic interactions and aromatic rings, which W112 can also provide. Y194 and F194 are similarly related. A reverse BLAST search gave the previous gene as the top hit and then this gene as the second hit. It is therefore concluded that this gene codes for aromatic aminotransferase, but with less confidence than the first gene.

Gene OID: 2505170441 EC: 2.6.1.1

This gene is predicted to also code for aromatic aminotransferase. A NCBI BLAST search gave other aminotransferases as hits [[3]](http://198.128.25.103/genome/converter.php#ref-3) with e-values ranging from 4E-58 to 4E-12 and around 25% to 30% identity. A PDB BLAST search gave an article on aromatic aminotransferase as a top hit [[19]](http://198.128.25.103/genome/converter.php#ref-19) with an e-value of 9.86089E-48 and 33% identity. Out of the active sites mentioned, 9 out of 14 of them matched. The mismatches were A96S, F121Y, V201A, M260D and T264I. Out of these mismatches, two of them were close mismatches, A96S and F121Y, because the mismatched amino acids are similar in structure. For the other 3, the article was searched to see if it mentioned these sites. V201 was mentioned in the article to be replaced by A201 in some cases. M260 was important as part of a hydrophobic pocket. D260 however, is not hydrophobic. It could be the case that D260 could be a single hydrophilic component of the pocket, while the rest of the pocket is overall hydrophobic. T264 is mentioned as important for hydrogen bonding with its hydroxyl group on its side chain, which I264 does not have. Also, a reverse BLAST search did not give this gene as the top hit, but rather the first gene. Despite that, this gene is still predicted to code for aromatic aminotransferase.

Gene OID: 2505169947

This gene is not predicted to code for an aromatic aminotransferase. All the top hits from NCBI BLAST were to LL-diaminopimelate aminotransferases {{NCBI: Q3AC10}} and a PDB BLAST search gave hits to N-Succinyldiaminopimelate aminotransferase [[29]](http://198.128.25.103/genome/converter.php#ref-29).

Gene OID: 2505167581

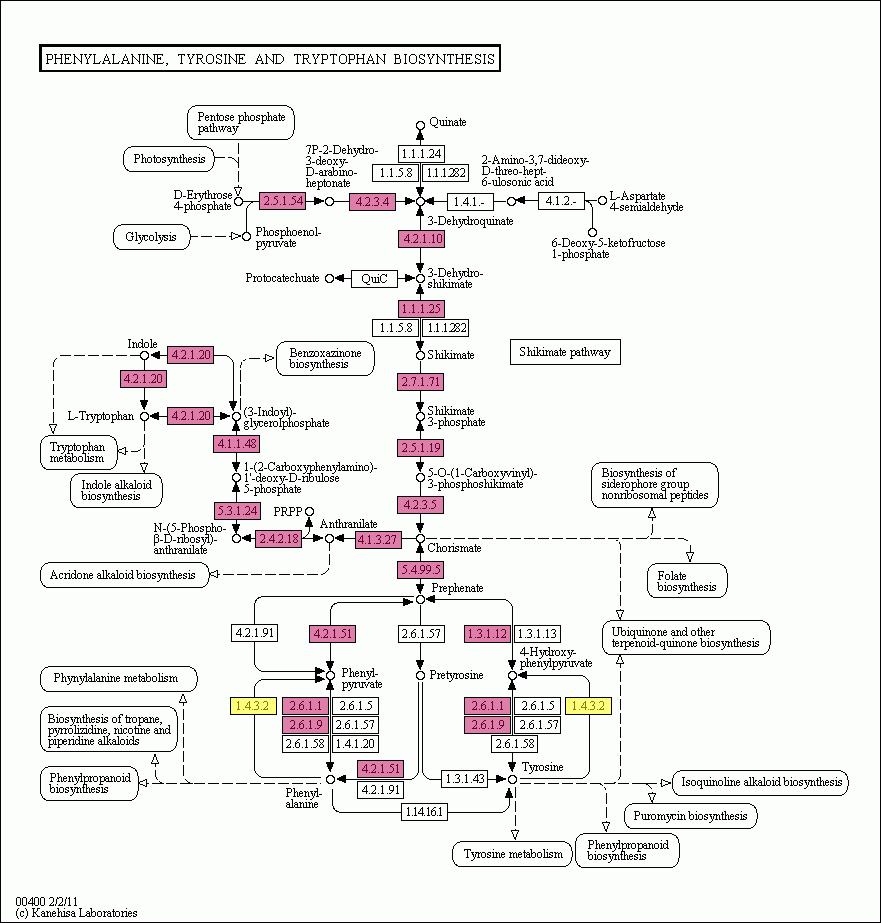
This gene is also not predicted to code for an aromatic aminotransferase. All the top hits from NCBI BLAST were to LL-diaminopimelate aminotransferases {{NCBI: Q3Z8H5}} and a PDB BLAST search gave hits to N-Succinyldiaminopimelate aminotransferase (PDB: 2DOU).

**Prephenate dehydrogenase**

Gene OID: 2505170102 EC: 1.3.1.12

This gene was predicted to code for prephenate dehydrogenase. A NCBI BLAST search gave as top hits prephenate dehydrogenases [[7]](http://198.128.25.103/genome/converter.php#ref-7) with e-values from 3E-31 to 3E-23 with around 30% identity. A PDB BLAST search gave an article on prephenate dehydrogenase crystal structure [[27]](http://198.128.25.103/genome/converter.php#ref-27) with an e-value of 1.62831E-27 and 29% identity. Out of the active sites mentioned, 7 out of 9 of them matched. The mismatches were E153A and H217L. E153 is mentioned as important for its polarity and the fact that it is charged. A153 possesses neither of these qualities. At site 217, H217 was important for its ability to form hydrogen bonds with its side chain, something L217 cannot do. As another check, a multiple sequence alignment was done on T-Coffee with unreviewed cyanobacterial FASTA sequences from UniProt. This showed that at site 153, some cyanobacteria also had the E153, but there were others that had the A153 residue. At site 217, two other cyanobacteria also had the H217 residue, but these looked like unreliable hits. There was one more reliable hit of a cyanobacteria with the L217 residue. Most of them had the V217 residue, which is similar to L217. The alignment shows that although the active sites don’t match, the mismatches could be found in other cyanobacterial FASTA sequences too. A reverse BLAST search also gives this gene as the top hit, so it is concluded that the gene codes for prephenate dehydrogenase.

**Comparison with Other Genomes**





S Fig. 3: KEGG pathway map of F, Y, W Amino Acid Biosynthesis for all cyanobacteria in IMG/ER

More than 75% of cyanobacteria are predicted by IMG/ER to have the genes necessary to undergo F, Y, and W amino acid biosynthesis. There were no new genes essential for the pathway that were not found in *Microcoleus vaginatus*, and the genes found in *Microcoleus vaginatus* were found in most other cyanobacteria. Cyanobacteria would be predicted to be able to synthesize these amino acids on their own because they cannot ingest amino acids by eating other organisms.

**Functional Studies**

Most of the genes in this pathway were easily annotated and were strongly predicted to code for their respective proteins. There were some problem genes however, that need to be further researched in the lab.

Gene 2505168075 was automatically annotated by IMG/ER and was predicted to code for anthranilate phosphoribosyltransferase. However, poor active site matching made this annotation questionable. This gene is probably misannotated. An experiment should be done to check that this gene really does not code for anthranilate phosphoribosyltransferase, by knocking out other genes that code for anthranilate phosphoribosyltransferase and seeing if it is still expressed.

Gene 2505167663 needs research done on it as well. This gene is predicted to code for chorismate mutase. The NCBI BLAST and PDB BLAST searches point to this fact, but only 6 out of 20 active sites matched. It would be good to double-check to see if this gene codes for chorismate mutase, or if the protein is expressed at all in *Microcoleus vaginatus*.

Genes 2505169488, 2505171077 and 2505170441 should also be checked. These genes were predicted to code for the aminotransferases that catalyze the final steps of F and Y biosynthesis. It should first be checked if they all code for aromatic aminotransferase or not. This protein is multifunctional and can catalyze the final step of both F and Y biosynthesis. It should also be seen if any of the genes code for aspartate aminotransferase, because this protein can only catalyze the final step of F biosynthesis. Different combinations of knockouts of the genes should be done, and it should be observed which proteins are expressed and which amino acids can be synthesized by *Microcoleus vaginatus*.

**Discussion**

Genes 2505166457 and gene 2505170171 are interesting because they both code for phospho-2-dehydro-3-deoxyheptonate aldolase, and with strong confidence. It would be interesting to see why this is the case, and to see if there is an evolutionary benefit to having two genes code for the same protein function, or if the different genes are expressed under different external circumstances.

Research should be done to see if *Microcoleus vaginatus* indeed synthesizes its own F and Y. If it does, it means that chorismate mutase and aromatic aminotransferase both exist in the organism, and it just becomes a matter of S Figuring out which genes code for which aminotransferases.

From the annotation, *Microcoleus vaginatus* is predicted to synthesize its own aromatic amino acids. As mentioned before, this makes sense, because it cannot eat other organisms and ingest amino acids. Therefore, it should synthesize its own aromatic amino acids.

**References**

[1] Ahn HJ, Yoon HJ, Lee B 2nd, Suh SW. "Crystal structure of chorismate synthase: a novel

FMN-binding protein fold and functional insights." J Mol Biol. Feb 27, 2004; 336(4):903-15.

[2] Chook YM, Gray JV, Ke H, Lipscomb WN. "The monofunctional chorismate mutase from

Bacillus subtilis. Structure determination of chorismate mutase and its complexes with a

transition state analog and prephenate, and implications for the mechanism of the

enzymatic reaction." J Mol Biol. Jul 29, 1994; 240(5):476-500.

[3] Cohen GN, Barbe V, Flament D, Galperin M, Heilig R, Lecompte O, Poch O, Prieur D,

Quérellou J, Ripp R, Thierry JC, Van der Oost J, Weissenbach J, Zivanovic Y, Forterre P.

"An integrated analysis of the genome of the hyperthermophilic archaeon Pyrococcus

abyssi." Mol Microbiol. 2003 Mar; 47(6):1495-512.

[4] Fucile G, Garcia C, Carlsson J, Sunnerhagen M, Christendat D. "Structural and biochemical

investigation of two Arabidopsis shikimate kinases: The heat-inducible isoform is

thermostable." Protein Sci. 2011 Jul; 20(7):1125-36.

[5] Goto M, Omi R, Miyahara I, Hosono A, Mizuguchi H, Hayashi H, Kagamiyama H, Hirotsu K.

"Crystal structures of glutamine:phenylpyruvate aminotransferase from Thermus

thermophilus HB8: induced fit and substrate recognition." J Biol Chem. Apr 16, 2004;

279(16):16518-25.

[6] Gray JV, Golinelli-Pimpaneau B, Knowles JR. "Monofunctional chorismate mutase from

Bacillus subtilis: purification of the protein, molecular cloning of the gene, and

overexpression of the gene product in Escherichia coli." Biochemistry. Jan 16, 1990;

29(2):376-83.

[7] Griffin HG, Gasson MJ. "Genetic aspects of aromatic amino acid biosynthesis in Lactococcus

lactis." Mol Gen Genet. Jan 6, 1995; 246(1):119-27.

[8] Henn-Sax M, Thoma R, Schmidt S, Hennig M, Kirschner K, Sterner R. "Two

(betaalpha)(8)-barrel enzymes of histidine and tryptophan biosynthesis have similar

reaction mechanisms and common strategies for protecting their labile substrates."

Biochemistry. Oct 8, 2002; 41(40):12032-42.

[9] Hennig M, Darimont BD, Jansonius JN, Kirschner K. "The catalytic mechanism of

indole-3-glycerol phosphate synthase: crystal structures of complexes of the enzyme

from Sulfolobus solfataricus with substrate analogue, substrate, and product." J Mol Biol.

Jun 7, 2002; 319(3):757-66.

[10] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[11] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo

K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M,

Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[12] Kim C, Xuong NH, Edwards S, Madhusudan, Yee MC, Spraggon G, Mills SE. "The crystal

structure of anthranilate phosphoribosyltransferase from the enterobacterium

Pectobacterium carotovorum." FEBS Lett. Jul 17, 2002; 523(1-3):239-46.

[13] Kim CW, Markiewicz P, Lee JJ, Schierle CF, Miller JH. "Studies of the hyperthermophile

Thermotoga maritima by random sequencing of cDNA and genomic libraries.

Identification and sequencing of the trpEG (D) operon." J Mol Biol. Jun 20, 1993;

231(4):960-81.

[14] Knöchel T, Ivens A, Hester G, Gonzalez A, Bauerle R, Wilmanns M, Kirschner K, Jansonius

JN. "The crystal structure of anthranilate synthase from Sulfolobus solfataricus:

functional implications." Proc Natl Acad Sci U S A. Aug 17, 1999; 96(17):9479-84.

[15] Lee CE, Goodfellow C, Javid-Majd F, Baker EN, Shaun Lott J. "The crystal structure of TrpD,

a metabolic enzyme essential for lung colonization by Mycobacterium tuberculosis, in

complex with its substrate phosphoribosylpyrophosphate." J Mol Biol. Jan 27, 2006;

355(4):784-97.

[16] Lee SJ, Ogasahara K, Ma J, Nishio K, Ishida M, Yamagata Y, Tsukihara T, Yutani K.

"Conformational Changes in the tryptophan synthase from a hyperthermophile upon

alpha2beta2 complex formation: crystal structure of the complex." Biochemistry. Aug 30,

2005; 44(34):11417-27.

[17] Liu JS, Cheng WC, Wang HJ, Chen YC, Wang WC. "Structure-based inhibitor discovery of

Helicobacter pylori dehydroquinate synthase." Biochem Biophys Res Commun. Aug 15,

2008; 373(1):1-7.

[18] Maes D, Gonzalez-Ramirez LA, Lopez-Jaramillo J, Yu B, De Bondt H, Zegers I, Afonina E,

Garcia-Ruiz JM, Gulnik S. "Structural study of the type II 3-dehydroquinate dehydratase

from Actinobacillus pleuropneumoniae." Acta Crystallogr D Biol Crystallogr. 2004 Mar;

60(Pt 3):463-71.

[19] Matsui I, Matsui E, Sakai Y, Kikuchi H, Kawarabayasi Y, Ura H, Kawaguchi S, Kuramitsu S,

Harata K. "The molecular structure of hyperthermostable aromatic aminotransferase with

novel substrate specificity from Pyrococcus horikoshii." J Biol Chem. Feb 18, 2000;

275(7):4871-9.

[20] Morollo AA, Eck MJ. "Structure of the cooperative allosteric anthranilate synthase from

Salmonella typhimurium." Nat Struct Biol. 2001 Mar; 8(3):243-7.

[21] Morris TR. "The clutch patterns of hens in constant illumination." Br Poult Sci. 1977 Jul;

18(4):397-405.

[22] Noback MA, Holsappel S, Kiewiet R, Terpstra P, Wambutt R, Wedler H, Venema G, Bron S.

"The 172 kb prkA-addAB region from 83 degrees to 97 degrees of the Bacillus subtilis

chromosome contains several dysfunctional genes, the glyB marker, many genes

encoding transporter proteins, and the ubiquitous hit gene." Microbiology. 1998 Apr; 144 (

Pt 4)():859-75.

[23] Padyana AK, Burley SK. "Crystal structure of shikimate 5-dehydrogenase (SDH) bound to

NADP: insights into function and evolution." Structure. 2003 Aug; 11(8):1005-13.

[24] Park H, Hilsenbeck JL, Kim HJ, Shuttleworth WA, Park YH, Evans JN, Kang C. "Structural

studies of Streptococcus pneumoniae EPSP synthase in unliganded state, tetrahedral

intermediate-bound state and S3P-GLP-bound state." Mol Microbiol. 2004 Feb;

51(4):963-71.

[25] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[26] Shumilin IA, Bauerle R, Wu J, Woodard RW, Kretsinger RH. "Crystal structure of the

reaction complex of 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase from

Thermotoga maritima refines the catalytic mechanism and indicates a new mechanism

of allosteric regulation." J Mol Biol. Aug 6, 2004; 341(2):455-66.

[27] Sun W, Singh S, Zhang R, Turnbull JL, Christendat D. "Crystal structure of prephenate

dehydrogenase from Aquifex aeolicus. Insights into the catalytic mechanism." J Biol

Chem. May 5, 2006; 281(18):12919-28.

[28] Tan K, Li H, Zhang R, Gu M, Clancy ST, Joachimiak A. "Structures of open (R) and close (T)

states of prephenate dehydratase (PDT)--implication of allosteric regulation by

L-phenylalanine." J Struct Biol. 2008 Apr; 162(1):94-107.

[29] Weyand S, Kefala G, Weiss MS. "The three-dimensional structure of

N-succinyldiaminopimelate aminotransferase from Mycobacterium tuberculosis." J Mol

Biol. Mar 30, 2007; 367(3):825-38.

[30] Wu J, Howe DL, Woodard RW. "Thermotoga maritima 3-deoxy-D-arabino-heptulosonate

7-phosphate (DAHP) synthase: the ancestral eubacterial DAHP synthase?" J Biol Chem.

Jul 25, 2003; 278(30):27525-31.

[31] Yamagata Y, Ogasahara K, Hioki Y, Lee SJ, Nakagawa A, Nakamura H, Ishida M, Kuramitsu

S, Yutani K. "Entropic stabilization of the tryptophan synthase alpha-subunit from a

hyperthermophile, Pyrococcus furiosus. X-ray analysis and calorimetry." J Biol Chem.

Apr 6, 2001; 276(14):11062-71.

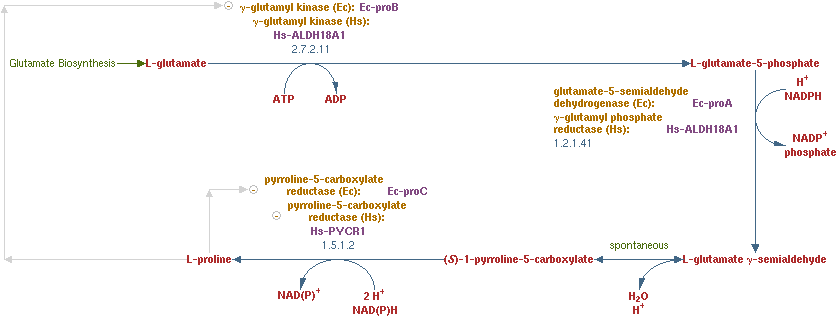
## Proline

Proline is one of the non-essential amino acids in many organisms, and as such many organisms can synthesize this amino acid on their own. These organisms all use, for the most part, the same pathway for proline biosynthesis, ultimately converting L-glutamate through a series of steps into L-proline.

This universal pathway is found in organisms ranging from bacteria to *Homo sapiens* , although the input to the pathway, L-glutamate, is formed by different means in different organisms. For example, in mammals, L-glutamate is formed in the mitochondrial matrix by an NAD(P)+-dependent glutamate dehydrogenase (EC 1.4.1.3) and transported to the cytoplasm, where the rest of the reactions occur, while in bacteria glutamate is formed by a NADP+-dependent enzyme (EC 1.4.1.4).

The first two enzymes, γ-glutamyl kinase and glutamate-5-semialdehyde dehydrogenase, form the γ-glutamyl kinase-GP-reductase multienzyme complex. Complex formation is necessary for the activity of the first enzyme, but not for the second. The product of the complex, L-glutamate γ-semialdehyde, rearranges spontaneously to form (*S*)-1-pyrroline-5-carboxylate, which is converted to L-proline in a single step, catalyzed by pyrroline-5-carboxylate reductase.

**Automated Annotation**



S Fig. 1: MetaCyc Pathway Map of Proline Amino Acid Biosynthesis

**Glutamate 5-kinase**

Gene OID: 2505170028 EC: 2.7.2.11

This gene is predicted to code for glutamate 5-kinase. A NCBI BLAST search gave only other glutamate 5-kinases as hits [[1]](http://198.128.28.118/genome/converter.php#ref-1) with e-values from 0 to 2E-90 and 45% to 75% identity. A PDB BLAST search gave as a top hit an article on the crystal structure of glutamate 5-kinase [[3]](http://198.128.28.118/genome/converter.php#ref-3) with an e-value of 1.91337E-77 and 45% identity. There were 16 active sites mentioned in this article, and out of those 16 active site, 15 of them matched. The lone mismatch was I53V, which isn’t a major mismatch because I53 and V53 are similar in structure. Additionally, a reverse BLAST search gave this gene as the top hit, so it the gene is strongly predicted to code for glutamate 5-kinase.

**Gamma-glutamyl phosphate reductase**

There were two genes predicted by IMG/ER to code for this protein. Both of them were predicted to code for gamma-glutamyl phosphate reductase.

Gene OID: 2505167847 EC: 1.2.1.41

This gene is predicted to code for gamma-glutamyl phosphate reductase. A NCBI BLAST search gave hits to gamma-glutamyl phosphate reductase except for the first hit [[5]](http://198.128.28.118/genome/converter.php#ref-5) with e-values from 2E-92 to 3E-62 and 31% to 38% identity. A PDB BLAST search gave as a top hit an article on gamma-glutamyl phosphate reductase crystal structure [[6]](http://198.128.28.118/genome/converter.php#ref-6) with an e-value of 9.25213E-55 and 30% identity. This article did not mention any active sites, and the next best PDB hit was to an unpublished article. Therefore, this gene is predicted to code for gamma-glutamyl phosphate reductase because NCBI and PDB BLAST searches give hits to gamma-glutamyl phosphate reductases.

Gene OID: 2505169340 EC: 1.2.1.41

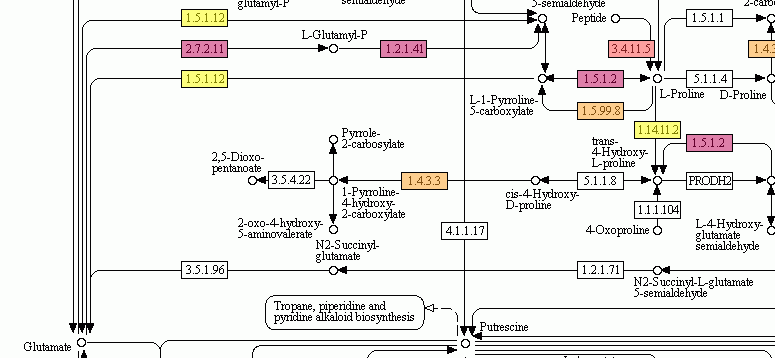
This gene is also predicted to code for gamma-glutamyl phosphate reductase, with higher confidence. A NCBI BLAST search gave hits mostly to gamma-glutamyl phosphate reductases {{NCBI: Q8YV15}} with e-values from 0 to 1E-114 and 43% to 76% identity. A PDB BLAST search gave as the top hit an unpublished article on gamma-glutamyl phosphate reductase (PDB: 1VLU) with an e-value of 6.02612E-102 and 46% identity. Since this article wasn’t published yet, there were no active sites that could be compared. The next best hit was to the same article as from before [[6]](http://198.128.28.118/genome/converter.php#ref-6) with an e-value of 2.29524E-93 and 42% identity, but as mentioned before, this article did not mention active sites. Therefore, like the previous gene, this gene is predicted to code for gamma-glutamyl phosphate reductase based on the NCBI and PDB BLAST hits, but this gene is more strongly predicted to code for the enzyme due to the greater strength of the BLAST hits.

**Pyrroline-5-carboxylate reductase**

Gene OID: 2505167806 EC: 1.5.1.2

This gene is predicted to code for pyrroline-5-carboxylate reductase. A NCBI BLAST search gave other pyrroline-5-carboxylate reductases as top hits [[2]](http://198.128.28.118/genome/converter.php#ref-2) with e-values ranging from 6E-112 to 2E-12 and around 24% to 60% identity. A PDB BLAST search gave as a top hit an article on the crystal structure of pyrroline-5-carboxylate reductase [[4]](http://198.128.28.118/genome/converter.php#ref-4) with an e-value of 1.73243E-45 and 39% identity. Out of 16 of the active sites mentioned in the article, 13 of them matched, with the 3 mismatches being G225A, H243A and N256E. The article did not specifically mention these residues as having specific catalytic function. Some judgements can be made about the mismatches though. The mismatch at site 225 is not major because G225 and A225 are very similar in structure, only differing in a side chain hydrogen versus a side chain methyl group. On the other hand, at site 243, H243 and A243 are quite different in structure, but H243 wasn’t mentioned for specific catalytic activity so this is not too much of a problem. For site 256, N256 and E256 are similar residues. Additionally, a reverse BLAST search gave this gene as the top hit, so it is predicted that the gene codes for pyrroline-5-carboxylate reductase.

**Comparison with Other Genomes**





S Fig. 2: KEGG Pathway Map of Proline Biosynthesis for comparison with other cyanobacterial genomes

Focusing in on the proline biosynthesis pathway that starts from glutamate in the bottom left corner, it can be seen that the 3 genes needed for proline biosynthesis are found in most other cyanobacteria as well. This is not surprising, as it was mentioned earlier that this pathway is a very universal one, so it would be expected that other cyanobacteria would use this pathway as well. The enzymes that *Microcoleus vaginatus* uses for proline biosynthesis are not unique or new.

**Functional Studies**

All of the enzymes required for the proline pathway were found to have genes coding for them. However, functional studies can be done on genes 2505167847 and 2505169340, the two genes predicted to code for gamma-glutamyl phosphate reductase. For both of these genes, PDB BLAST searches did not give hits to articles that mentioned article sites, so there is some uncertainty about the annotation because of this. It would be good to do some functional studies to double check and see that both these genes do in fact cod for gamma-glutamyl phosphate reductase.

Additionally, it is interesting to see two genes coding for proteins with the same function. Research should be done to see why this redundancy exists. It should be observed if having these two genes confers some sort of evolutionary benefit to the organism, or if having both of them is necessary for the organism to function. It should also be seen if the organism can function with just one of the genes expressed, and if the genes are expressed or repressed in reaction to different external conditions.

**Discussion**

From this annotation report, *Microcoleus vaginatus* is predicted to be able to synthesize its own proline. This matches up to experiment, as the bug was observed to be able to grow without being given amino acid supplements. *Microcoleus vaginatus* and most cyanobacteria have the enzymes necessary to undergo this pathway; this pathway is not unique to *Microcoleus vaginatus*.

**References**

[1] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[2] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N,

Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda

M, Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[3] Marco-Marín C, Gil-Ortiz F, Pérez-Arellano I, Cervera J, Fita I, Rubio V. "A novel two-domain

architecture within the amino acid kinase enzyme family revealed by the crystal structure

of Escherichia coli glutamate 5-kinase." J Mol Biol. Apr 13, 2007; 367(5):1431-46.

[4] Meng Z, Lou Z, Liu Z, Li M, Zhao X, Bartlam M, Rao Z. "Crystal structure of human

pyrroline-5-carboxylate reductase." J Mol Biol. Jun 23, 2006; 359(5):1364-77.

[5] Nakamura Y, Kaneko T, Sato S, Mimuro M, Miyashita H, Tsuchiya T, Sasamoto S, Watanabe

A, Kawashima K, Kishida Y, Kiyokawa C, Kohara M, Matsumoto M, Matsuno A,

Nakazaki N, Shimpo S, Takeuchi C, Yamada M, Tabata S. "Complete genome structure

of Gloeobacter violaceus PCC 7421, a cyanobacterium that lacks thylakoids." DNA Res.

Aug 31, 2003; 10(4):137-45.

[6] Page R, Nelson MS, von Delft F, Elsliger MA, Canaves JM, Brinen LS, Dai X, Deacon AM,

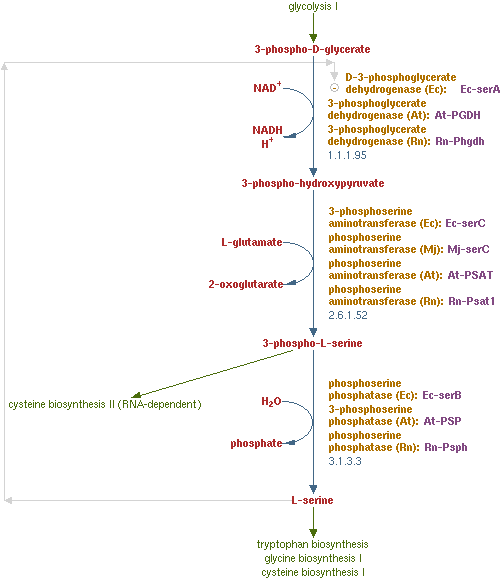
Floyd R, Godzik A, Grittini C, Grzechnik SK, Jaroszewski L, Klock HE, Koesema E, Kovarik JS, Kreusch A, Kuhn P, Lesley SA, McMullan D, McPhillips TM, Miller MD, Morse A, Moy K, Ouyang J, Robb A, Rodrigues K, Schwarzenbacher R, Spraggon G, Stevens RC, van den Bedem H, Velasquez J, Vincent J, Wang X, West B, Wolf G, Hodgson KO, Wooley J, Wilson IA. "Crystal structure of gamma-glutamyl phosphate reductase (TM0293) from Thermotoga maritima at 2.0 A resolution." Proteins. Jan 1, 2004; 54(1):157-61.

## Serine

Serine biosynthesis is a major metabolic pathway. Its end product, L-serine, is not only used in protein synthesis, but also as a precursor for the biosynthesis of glycine, cysteine, tryptophan, and phospholipids. In addition, it directly or indirectly serves as a source of one-carbon units for the biosynthesis of various compounds.

Regulation of the pathway is mainly accomplished by feedback inhibition of the enzyme that catalyzes the first committed step, D-3-phosphoglycerate dehydrogenase. The second enzyme in the pathway, 3-phosphoserine aminotransferase , requires pyridoxal 5'-phosphate (PLP) as a cofactor and is also required for the biosynthesis of PLP itself. Thus the cell must ensure that the supply of PLP is adequate. Little biochemical work has been done on the final enzyme of the pathway, phosphoserine phosphatase.

**Automated Annotation**



S Fig. 1: MetaCyc Pathway Map of serine amino acid biosynthesis

**D-3-phosphoglycerate dehydrogenase**

Gene OID: 2505167835 EC: 1.1.1.95

This gene is predicted to code for D-3-phosphoglycerate dehydrogenase. A NCBI BLAST search gave D-3-phosphoglycerate dehydrogenases as top hits [[7]](http://198.128.28.118/genome/converter.php#ref-7) with e-values from 0 to 7E-96 and 39% to 77% identity. A PDB BLAST search gave as a top hit an article on the crystal structure of D-3-phosphoglycerate dehydrogenase [[2]](http://198.128.28.118/genome/converter.php#ref-2) with an e-value of 5.18868E-98 and 40% identity. Out of the active sites mentioned in this article, 3 out out of 5 of them matched up. The two mismatches were Y461H and R464M. The Y461 residue was mentioned in the article, but it was not mentioned why it was important for catalytic function. Y461 and H461 are similar anyways, both having aromatic rings. At site 464, R464 was mentioned for its ability to hydrogen bond, which M464 cannot do. This discrepancy seems minor though, compared to the other evidence for the gene coding for the enzyme. Additionally, a reverse BLAST search gave this gene as the top hit, so it is predicted that the gene codes for D-3-phosphoglycerate dehydrogenase.

**Phosphoserine aminotransferase**

Gene OID: 2505166538 EC: 2.6.1.52

There was no gene predicted by IMG/ER to code for this protein, so a search was done on UniProt for cyanobacterial phosphoserine aminotransferase FASTA sequences. These sequences were BLASTed against the *Microcoleus vaginatus* genome, and gene 2505166538 was one of the hits. With this gene, a NCBI BLAST search was done. This search gave phosphoserine aminotransferases as hits [[1]](http://198.128.28.118/genome/converter.php#ref-1), but with rather high e-values. The best hit had an e-value of 2.3 and 35% identity, while the worst hit had an e-value of 9.8 and 33% identity. A PDB BLAST search was also done, and this search gave an article on phosphoserine aminotransferase as a top hit [[5]](http://198.128.28.118/genome/converter.php#ref-5), but with a high e-value of 1.61642 and 37% identity. Furthermore, out of 12 of the active sites mentioned in the article, only 3 of the active site residues matched. There were deletions at sites 102, 174 and 177. The rest of the sites were mismatches, with the mismatches being R42W, T153L, N239G, T240N, H328L, and R335Q. This data sheds doubt on if this gene actually codes for phosphoserine aminotransferase, but no other genes were found that could code for the enzyme, so this could be the gene that codes for the enzyme. More research needs to be done on this gene for any conclusive decision can be made about whether or not the enzyme is expressed, and if this gene is the the gene that codes for the enzyme.

**Phosphoserine phosphatase**

EC: 3.1.3.3

There was a gene predicted by IMG/ER to code for this enzyme, gene 2505170791. This gene was actually predicted to not code for phosphoserine phosphatase. A search was done for other cyanobacterial FASTA sequences for this enzyme, and these FASTA sequences were BLASTed against the *Microcoleus vaginatus* genome in order to find hits. This gave 8 other genes as possible hits for phosphoserine phosphatase. All these hits are putative and need to be double-checked by experiment.

Gene OID: 2505170791

This gene is predicted to not code for phosphoserine phosphatase. A NCBI BLAST search gave no hits to phosphoserine phosphatase, and a PDB BLAST also did not give hits to phosphoserine phosphatase. Therefore, this gene is predicted to not code for the enzyme.

Gene OID: 2505166541

This gene could code for phosphoserine phosphatase. A NCBI BLAST search gave a hit to Phosphoserine phosphatase rsbU [[12]](http://198.128.28.118/genome/converter.php#ref-12) with an e-value of 8E-20 and 32% identity. No hits were found from a PDB BLAST hit. This lack of data makes it hard to say whether or not the gene actually codes for phosphoserine phosphatase. More research needs to be done.

Gene OID: 2505167639

This gene could code for phosphoserine phosphatase. A NCBI BLAST search gave a hit to Phosphoserine phosphatase rsbU [[12]](http://198.128.28.118/genome/converter.php#ref-12) with an e-value of 5E-10 and 25% identity. No hits were found from a PDB BLAST hit. This lack of data makes it hard to say whether or not the gene actually codes for phosphoserine phosphatase. More research needs to be done.

Gene OID: 2505167603

This gene could code for phosphoserine phosphatase. A NCBI BLAST search gave a hit to Phosphoserine phosphatase rsbU [[12]](http://198.128.28.118/genome/converter.php#ref-12) with an e-value of 5E-15 and 27% identity. No hits were found from a PDB BLAST hit. This lack of data makes it hard to say whether or not the gene actually codes for phosphoserine phosphatase. More research needs to be done.

Gene OID: 2505169736

This gene could code for phosphoserine phosphatase. A NCBI BLAST search gave a hit to Phosphoserine phosphatase rsbU [[12]](http://198.128.28.118/genome/converter.php#ref-12) with an e-value of 4E-16 and 27% identity. No hits were found from a PDB BLAST hit. This lack of data makes it hard to say whether or not the gene actually codes for phosphoserine phosphatase. More research needs to be done.

Gene OID: 2505168969

This gene could code for phosphoserine phosphatase. A NCBI BLAST search gave a hit to Phosphoserine phosphatase rsbU [[12]](http://198.128.28.118/genome/converter.php#ref-12) with an e-value of 7E-27 and 27% identity. No hits were found from a PDB BLAST hit. This lack of data makes it hard to say whether or not the gene actually codes for phosphoserine phosphatase. More research needs to be done.

Gene OID: 2505167281

This gene could code for phosphoserine phosphatase. A NCBI BLAST search gave a hit to Phosphoserine phosphatase rsbU [[12]](http://198.128.28.118/genome/converter.php#ref-12) with an e-value of 7E-14 and 26% identity. No hits were found from a PDB BLAST hit. This lack of data makes it hard to say whether or not the gene actually codes for phosphoserine phosphatase. More research needs to be done.

Gene OID: 2505168773

This gene could code for phosphoserine phosphatase. A NCBI BLAST search gave a hit to Phosphoserine phosphatase rsbU [[12]](http://198.128.28.118/genome/converter.php#ref-12) with an e-value of 2E-12 and 25% identity. No hits were found from a PDB BLAST hit. This lack of data makes it hard to say whether or not the gene actually codes for phosphoserine phosphatase. More research needs to be done.

Gene OID: 2505165951

This gene could code for phosphoserine phosphatase. A NCBI BLAST search gave a hit to phosphoserine phosphatase rsbU [[12]](http://198.128.28.118/genome/converter.php#ref-12) with an e-value of 2E-11 and 27% identity. No hits were found from a PDB BLAST hit. This lack of data makes it hard to say whether or not the gene actually codes for phosphoserine phosphatase. More research needs to be done.

**Serine hydroxymethyltransferase**

Gene OID: 2505168407 EC: 2.1.2.1

This gene is predicted to code for serine hydroxymethyltransferase. A NCBI BLAST search gave only serine hydroxymethyltransferases as hits [[6]](http://198.128.28.118/genome/converter.php#ref-6) with e-values of 0 and 62% to 86% identity. A PDB BLAST search gave as a top hit an article on serine hydroxymethyltransferase [[11]](http://198.128.28.118/genome/converter.php#ref-11) with an e-value of 1.56754E-142 and 61% identity. This article mentioned active sites, and out of these mentioned active sites, 21 out of 22 of them matched. The sole mismatch was R58K. The R58 residue was not mentioned in the article as having specific catalytic funciton, and the R58 and K58 residues are very similar, both having long carbon backbones with nitrogens on them as side groups. Additionally, a reverse BLAST search gave this gene as the top hit, so this gene is strongly predicted to code for serine hydroxymethyltransferase.

**Serine dehydratase**

EC: 4.3.1.19

Two genes were predicted by IMG/ER to code for this enzyme. The first gene, 2505166192, was predicted to code for serine dehydratase. The second gene, 2505167467, was predicted to code for a bifunctional protein with serine dehydratase and serine racemase functionality.

Gene OID: 2505166192

This gene is predicted to code for serine dehydratase. A NCBI BLAST search gave as top hits threonine dehydratases, which are the same as serine dehydratases, [[8]](http://198.128.28.118/genome/converter.php#ref-8). These hits had e-values around 0 to 8E-17, and from around 26% to 69% identity. A PDB BLAST search gave as a top hit an article on the enzyme threonine deanimase, which is bifunctional for serine dehydratase [[3]](http://198.128.28.118/genome/converter.php#ref-3). This hit had an e-value of 9.10007E-145 and 51% identity. Out of the active sites mentioned in the article, 19 out of 30 of them matched. The 11 mismatches were R55M, H59F, L171M, Q175R, A178Q, S315C, H322D, Y326F, M414L, Y421H and F510L. Only 3 of these active site residues were mentioned in the article. L171 is mentioned for being part of a hydrophobic region, and M171 is hydrophobic too. Q175 is mentioned for being able to form multiple hydrogen bonds, and so can R175. A178 contributes to a hydrophobic region, but Q178 is not hydrophobic. S315 is mentioned for being able to hydrogen bond with its side chain, but C315 cannot. To further shed light on the mismatches, a multiple sequence alignment was done on T-Coffee with cyanobacterial FASTA sequences from UniProt.

At site 55, nor R55 residues were found, but some M55 residues were found. At site 59, no H59 residues were found, and most of the residues were F59. For site 171, both L171 and M171 residues were found, and at site 175, both Q175 and R175 residues were found. At site 178, no A178 residues were found, but some Q178 residues were. At site 315, there were some S315 residues, but mostly C315 residues. At site 322, no H322 residues were found, but some D322 ones were. At site 326, only one Y326 residue was found and all other residues were F326. At site 414, no M414 residues were found, but almost all of the residues were L414. For site 421, no Y421 residues were found, but all of the residues were H421. Finally, at site 510, there were some F510 residues, but the majority of the residues were L510.

The multiple sequence alignment showed that that mismatches were justifiable since they were seen in other cyanobacteria. Additionally, a reverse BLAST search gave this gene as the top hit, so this gene is predicted to code for serine dehydratase.

Gene OID: 2505167467 EC: 5.1.1.18/ 4.3.1.19

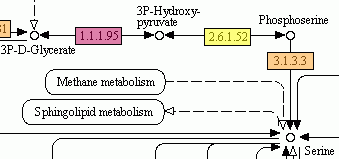
This gene could code for serine racemase or serine dehydratase. A NCBI BLAST search was done and gave as the top hits serine racemases [[13]](http://198.128.28.118/genome/converter.php#ref-13) with e-values from 2E-110 to 4E-73 and 44% to 54% identity. A PDB BLAST search gave as a top hit an article on serine racemase crystal structure [[4]](http://198.128.28.118/genome/converter.php#ref-4) with an e-value of 6.22249E-93 and 53% identity. 23 out of 24 of the active sites mentioned in the article matched, with the lone mismatch being S33E. These two residues are similar in structure, with E33 having an extra double-bonded oxygen. This gene is therefore predicted to code for serine racemase.

It was also checked to see if the gene could code for serine dehydratase as well. The NCBI BLAST search also gave many threonine dehydratases hits (same as serine dehydratase) [[9]](http://198.128.28.118/genome/converter.php#ref-9) with e-values from 9E-71 to 1E-12 and 30% to 42% identity. The PDB BLAST search also gave an article on serine dehydratase crystal structure [[10]](http://198.128.28.118/genome/converter.php#ref-10) with an e-value of 1.22287E-64 and 43% identity. Out of the active sites mentioned in the article, 26 out of 33 of the active site residues matched. The 7 active site mismatches were N34E, Y35T, F63C, D119G, P182C, I189L and Q275E. The I189L mismatch is not much of a mismatch, since I189 and L189 are so similar in structure. Three of the mismatches were mentioned in the article. For N34, it was mentioned as important for hydrogen bonding with its side chain nitrogen. E34 doesn’t have a side chain nitrogen, but it does have a side chain oxygen it can hydrogen bond with. For site 119, D119 was mentioned as being able to hydrogen bond with its side chain oxygen. G119 has no such ability to hydrogen bond. For site Q275, this residue was mentioned to hydrogen bond with its side chain oxygen. E275 also has such a side chain oxygen to hydrogen bond.

The rest of the mismatches not mentioned in the article were checked by a multiple sequence alignment on T-Coffee against other cyanobacterial FASTA sequences from UniProt. This showed that at site 35, no other cyanobacteria had the Y35 or T35 residue, but most of them had the N35 residue. At site 63, most other cyanobacteria had the Y63 residue, and a few had the F63 residue. None of them had the C63 residue. Site 119 doesn’t seem to be a strongly conserved residue site, because there is a wide variety of residues at this spot. There is D119 and also G119 among other cyanobacteria. At site 182, most of the residues were A182, with one other C182 residue. There were no P182 residues.

From the multiple sequence alignment, it seems that the site 63 mismatch, F63C is the only major mismatch that cannot be supported. Despite this, the rest of the evidence suggests that there is still a strong possiblity that the gene could code for serine dehydratase. It seems that this gene codes for a bifunctional protein.

**Comparison with Other Genomes**





S Fig. 2: KEGG Pathway Map of serine amino acid biosynthesis for comparison with other cyanobacteria

Focusing on the serine biosynthesis pathway that starts with 3P-D-Glycerate, the comparison with other cyanobacterial genomes supported the findings for *Microcoleus vaginatus*. The first enzyme, EC 1.1.1.95, was predicted to be in *Microcoleus vaginatus* and it is also predicted to be in most cyanobacteria. The next two enzymes, EC 2.6.1.52 and 3.1.3.3, were not conclusively found in *Microcoleus vaginatus*; they may or may not be there. For other cyanobacteria, these two enzymes were much less common, and so it would not be unusual if *Microcoleus vaginatus* is missing these enzymes. However, this finding begs the question of how cyanobacteria can synthesize serine, or if they do synthesize serine. Assuming that they do synthesize serine, there must be an alternate pathway or some not-yet discovered enzyme that allows for serine biosynthesis.

**Functional Studies**

Functional studies need to be done on the phosphoserine aminotransferase and phosphoserine phosphatase enzymes, and the genes predicted to potentially code for these enzymes.

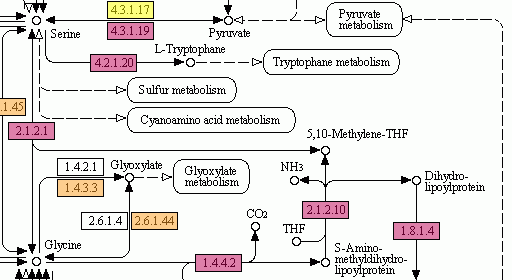
Functional studies should be done on gene 2505166538. This gene was predicted to have a function by IMG/ER, but a NCBI BLAST search and PDB BLAST search gave hits to phosphoserine aminotransferases, although with high e-values. Experiments should be done to see if the phosphoserine aminotransferase enzyme (EC 2.6.1.52) is found in *Microcoleus vaginatus*; if it is, this gene is the top candidate for expressing that enzyme.

Similarly, functional studies need to be done on genes 2505166541, 2505167639, 2505167603, 2505169736, 2505168969, 2505167281, 2505168773 and 2505165951, which each could potentially code for phosphoserine phosphatase (EC 3.1.3.3). First of all, an experiment should be done to establish whether or not phosphoserine phosphatase is found in *Micrcoleus vaginatus*. If it is, then functional studies need to be done to S Figure out which gene or genes code for phosphoserine phosphatase. This would need to be done by growing mutants with different genes repressed and seeing if phosphoserine phosphatase is still found in the organism.

Finally, functional studies should be done on gene 2505167467. This gene is predicted to code for a bifunctional protein with serine racemase and serine dehydratase function. A study should be done to confirm or disprove this prediction. Studies should also be done to see why there are multiple serine dehydratases in the organism, and if so, why this is the case.

**Discussion**

From the annotation, it is unclear whether or not *Microcoleus vaginatus* can synthesize serine using this pathway. It is known from experiment, however, that *Microcoleus vaginatus* indeed does synthesize its own serine, because it has been observed to grow without being given amino acids. If *Microcoleus vaginatus* does not synthesize serine by this pathway, then there must be some alternate pathway or some different enzyme that it can use to synthesize serine. A search for an alternate pathway was done on KEGG by looking at a pathway map and seeing if there were alternate ways to arrive at serine as a product. The KEGG map is shown below.





S Fig. 3: KEGG Pathway Map of alternate routes for serine biosynthesis

Looking at S Figure 3, two pathways are shown that lead to serine biosynthesis. The first is the pathway from glycine and 5,10-methylene-THF via enzyme 2.1.2.1. The second is the pathway from pyruvate by enzyme 4.3.1.19. *Micrcoleus vaginatus* has the genes to undergo both of these pathways, so it uses these two alternate pathways for serine biosynthesis, instead of the pathway shown in S Figure 1. Studies should be done to determine if both pathways or only one of them is used. Additionally, studies should be done to see if one pathway is preferred over the other, and whether that is due to chance, evolutionary selection or external conditions.

**References**

[1] C. elegans Sequencing Consortium. "Genome sequence of the nematode C. elegans: a

platform for investigating biology." Science. Dec 11, 1998; 282(5396):2012-8.

[2] Dey S, Grant GA, Sacchettini JC. "Crystal structure of Mycobacterium tuberculosis

D-3-phosphoglycerate dehydrogenase: extreme asymmetry in a tetramer of identical

subunits." J Biol Chem. Apr 15, 2005; 280(15):14892-9.

[3] Gallagher DT, Gilliland GL, Xiao G, Zondlo J, Fisher KE, Chinchilla D, Eisenstein E.

"Structure and control of pyridoxal phosphate dependent allosteric threonine

deaminase." Structure. Apr 15, 1998; 6(4):465-75.

[4] Goto M, Yamauchi T, Kamiya N, Miyahara I, Yoshimura T, Mihara H, Kurihara T, Hirotsu K,

Esaki N. "Crystal structure of a homolog of mammalian serine racemase from

Schizosaccharomyces pombe." J Biol Chem. Sep 18, 2009; 284(38):25944-52.

[5] Hester G, Stark W, Moser M, Kallen J, Marković-Housley Z, Jansonius JN. "Crystal structure

of phosphoserine aminotransferase from Escherichia coli at 2.3 A resolution: comparison

of the unligated enzyme and a complex with alpha-methyl-l-glutamate." J Mol Biol. Feb

26, 1999; 286(3):829-50.

[6] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[7] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N,

Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda

M, Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[8] May BJ, Zhang Q, Li LL, Paustian ML, Whittam TS, Kapur V. "Complete genomic sequence

of Pasteurella multocida, Pm70." Proc Natl Acad Sci U S A. Mar 13, 2001; 98(6):3460-5.

[9] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[10] Simanshu DK, Savithri HS, Murthy MR. "Crystal structures of Salmonella typhimurium

biodegradative threonine deaminase and its complex with CMP provide structural

insights into ligand-induced oligomerization and enzyme activation." J Biol Chem. Dec

22, 2006; 281(51):39630-41.

[11] Trivedi V, Gupta A, Jala VR, Saravanan P, Rao GS, Rao NA, Savithri HS, Subramanya HS.

"Crystal structure of binary and ternary complexes of serine hydroxymethyltransferase

from Bacillus stearothermophilus: insights into the catalytic mechanism." J Biol Chem.

May 10, 2002; 277(19):17161-9.

[12] Voelker U, Dufour A, Haldenwang WG. "The Bacillus subtilis rsbU gene product is

necessary for RsbX-dependent regulation of sigma B." J Bacteriol. 1995 Jan;

177(1):114-22.

[13] Yamauchi T, Goto M, Wu HY, Uo T, Yoshimura T, Mihara H, Kurihara T, Miyahara I, Hirotsu

K, Esaki N. "Serine racemase with catalytically active lysinoalanyl residue." J Biochem.

2009 Apr; 145(4):421-4.

# Fermentation

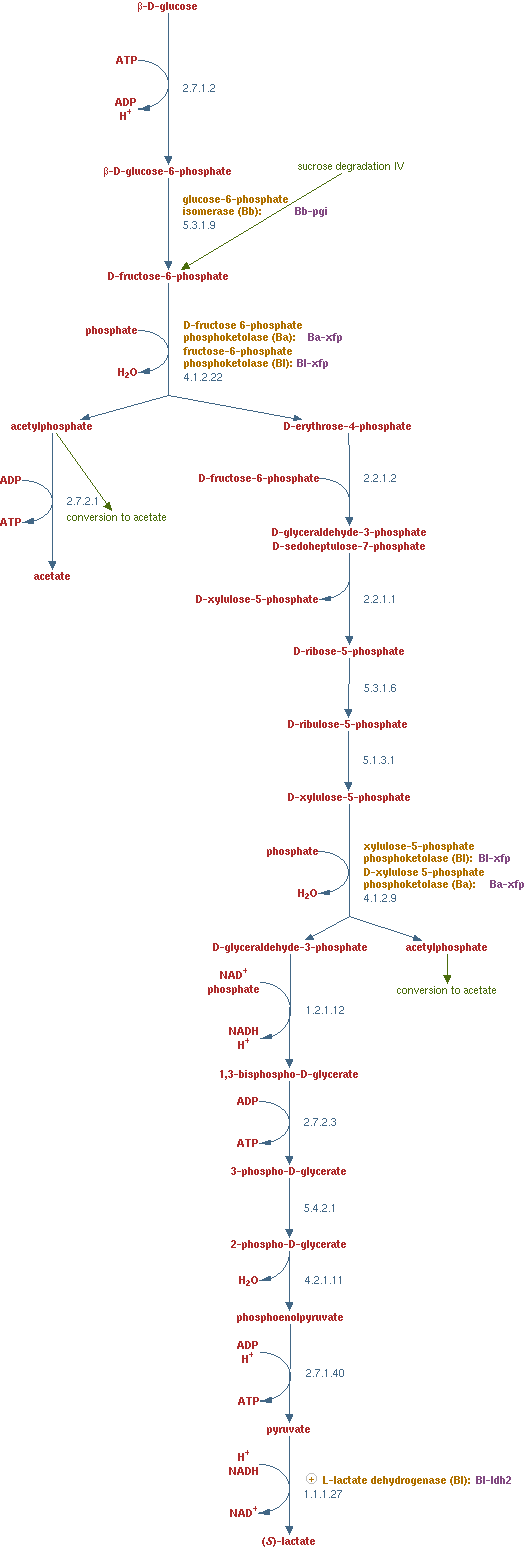
## Bifidobacterium Shunt

Cyanobacteria are documented to undergo fermentation [[13]](http://198.128.26.117/genome/converter.php#ref-13) via different fermentation pathways. One such pathway that *Microcoleus vaginatus* looked like it could go through was the Bifidobacterium Shunt, which is a fermentation pathway.

This fermentation pathway, which is known as the "Bifidobacterium shunt" or the "fructose-6-phosphate pathway" yields 3 mols of acetate and 2 mols of lactate for 2 mols of glucose, with production of 5 mols of ATP. The key enzyme in the pathway is xylulose-5-phosphate phosphoketolase / fructose-6-phosphate phosphoketolase, which catalyzes two important steps: splitting D-fructose-6-phosphate into D-erythrose-4-phosphate and acetylphosphate, and splitting D-xylulose-5-phosphate into D-glyceraldehyde-3-phosphate and acetylphosphate. This enzyme has often been used as a tool in the identification of Bifidobacteria, even though such enzymes have been found in other organisms.

Acetylphosphate, which is formed in both phosphoketolase-catalyzed steps, is converted to acetate. D-erythrose-4-phosphate is processed along with a second D-fructose-6-phosphate molecule by enzymes of the pentose phosphate pathway to D-xylulose-5-phosphate. This compound is then split in the second reaction catalyzed by phosphoketolase to D-glyceraldehyde-3-phosphate and acetylphosphate. D-glyceraldehyde-3-phosphate is processed via the glycolytic Embden-Meyerhof pathway to pyruvate, which is subsequently converted to (*S*)-lactate.

**Automated Annotation**



S Fig.1: MetaCyc pathway map of the bifidobacterium shunt. Full image at http://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=P124-PWY&detail-level=2&detail-level=1

**Acetate Kinase**

Gene OID: 2505165972 EC: 2.7.2.1

The gene is predicted to code for acetate kinase. A NCBI BLAST search gave as top hits other acetate kinases [[8]](http://198.128.26.117/genome/converter.php#ref-8) as top hits, with e-values as low as 0 and percent identities as high as 68%. A PDB BLAST search gave articles on acetate kinases as top hits, with the best hit having an e-value of 4.72107E-85 and 42% identity [[3]](http://198.128.26.117/genome/converter.php#ref-3). 17 active sites were mentioned in the article, and all 17 were matched. A reverse BLAST search gives this gene as the first hit. This gene is therefore strongly predicted to code for acetate kinase.

**Fructose-6-Phosphate Phosphoketolase**

Gene OID: 2505165968 EC: 4.1.2.22

The gene is predicted to code for fructose-6-phosphate phosphoketolase (F6P phosphoketolase). A NCBI BLAST gave as top hits probable phosphoketolases [[16]](http://198.128.26.117/genome/converter.php#ref-16) with e-values around 0 and around 75% identity. A PDB BLAST search gave an article on phosphoketolase [[21]](http://198.128.26.117/genome/converter.php#ref-21) with an e-value of 0 and 51% identity. Active sites were mentioned in this article. 19 out of 19 active sites matched, so this gene is very likely coding for F6P phosphoketolase.

**L-Lactate Dehdyrogenase**

Gene OID: 2505168756 EC: 1.1.1.27/1.1.1.32

This gene could code for L-lactate dehydrogenase. This gene was previously predicted to code for malate dehydrogenase, but there is a possibility it could code for a bifunctional gene. A NCBI BLAST search gave as top hits malate dehydrogenases, but there is one hit for L-lactate dehydrogenase [[12]](http://198.128.26.117/genome/converter.php#ref-12) with an e-value of 6E-96 and 47% identity. A PDB BLAST search also mainly gave hits to articles on malate dehydrogenase, but some hits were articles for L-lactate dehydrogenase, like a hit for an article on L-lactate dehdyrogenase in *Lactobacillus casei* [[1]](http://198.128.26.117/genome/converter.php#ref-1) with an e-value of 3.05424E-41 and 29% identity. A reverse BLAST hit using the FASTA sequence from this hit gave gene 2505168756 as the top and only hit. Therefore, it is concluded that gene 2505168756 might code for a protein bifunctional for EC 1.1.1.27 and 1.1.1.32 function.

**Glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)**

Gene OID: 2505170636 EC 1.2.1.12

This gene is predicted to code for the Glyceraldehyde-3-phosphate dehydrogenase enzyme (EC 1.2.1.12). An NCBI BLAST search gave other Glyceraldehyde-3-phosphate dehydrogenase enzymes as top hits, with an e-value of 0 and a query coverage of 99% (NCBI: P34916). A reverse BLAST search also gave gene 2505170636 as a top hit, with an e-value of 0. A PDB BLAST search gave other Glyceraldehyde-3-phosphate dehydrogenase enzymes as top hits. The active sites were compared to the Glyceraldehyde-3-phosphate dehydrogenase enzyme in E.coli, and there was a 100% active site match. This particular hit had an e-value of 2.15146E-118 and an identity of 63% (PDB: 1S7C).

**Transketolase**

Gene OID: 2505170664 EC 2.2.1.1

This gene was predicted to code for the Transketolase enzyme (EC 2.2.1.1). An NCBI BLAST search gave other Transketolase enzymes as top hits, with an e-value of 0 and a query coverage of 100%. [[7]](http://198.128.26.117/genome/converter.php#ref-7) A reverse BLAST search also gave gene 2505170664 as a top hit, with an e-value of 0. A PDB BLAST search gave other Transketolase enzymes as top hits. The top PDB hit was the Transketolase enzyme structure within yeast. This particular hit had an e-value of 0 and an identity of 51%. Additionally, there was a 100% active site match. [[26]](http://198.128.26.117/genome/converter.php#ref-26)

**Transaldolase**

Gene OID: 2505170982 and 2505169096 EC: 2.2.1.2

There were two different genes that showed up as hits for this enzyme. Gene 2505170982 was predicted to be a fructose-6-phosphate aldolase while gene 2505169096 was predicted to be a transaldolase.

Gene 2505170982 was predicted to be a fructose-6-phosphate aldolase on the basis of an article from a PDB BLAST search [[23]](http://198.128.26.117/genome/converter.php#ref-23). This article had a list of active site residues that were conserved for both aldolases and transaldolases. All these active sites matched with the ones in *Microcoleus vaginatus*. The article also had a list of active site residues that were different between aldolases and transaldolases. Out of these active sites, 5/8 sites matched the sites for an aldolase. No sites matched the sites for a transaldolase. The other 3 sites were mismatches. A mismatch at site 107, L to A, could be explained by the fact that these two amino acids are similar in structure. At site 57, F was replaced by Y, and they are similar except for a hydroxyl group on the aromatic ring. Finally, at site 129, A was replaced by I, and they are similar in that they are both carbon based side chains, only of different lengths. From this data, it was concluded that the gene was incorrectly annotated and should actually be a fructose-6-phosphate aldolase. Also, a NCBI BLAST search gave as top hits both transaldolases and fructose-6-phosphate aldolases [[18]](http://198.128.26.117/genome/converter.php#ref-18) so the results from the PDB article were the deciding factor.

Gene 2505169096 was predicted to be a transaldolase on the basis of NCBI and PDB BLASTs along with an InterProScan search. InterProScan predicted the gene to code for a transaldolase. NCBI BLAST gave other transaldolases as top hits, and PDB BLAST brought up articles about transaldolases as top hits [[20]](http://198.128.26.117/genome/converter.php#ref-20).

**Glucokinase**

Gene OID: 2505166819 EC: 2.7.1.2

This gene may code for a glucokinase, but it is unclear. A NCBI BLAST search gave as top hits other glucokinases with e-values between 5E-43 and 5E-39 and around 35% identity, but all the hits after these top three hits were not for glucokinases [[22]](http://198.128.26.117/genome/converter.php#ref-22). A reverse BLAST search gave the gene as a top hit, with an e-value of 4e-23. A PDB BLAST search gave as the top hit an unpublished article on glucokinase (PDB: 2QM1) with an e-value of 1.98516E-47 and 36%, but all the hits after that one were for other proteins besides glucokinases. Therefore, it is unclear if these gene actually codes for a glucokinase or not. Further research needs to be done.

Gene OID: 2505170243 EC: 2.7.1.2

This gene was predicted to code for glucokinase. A NCBI BLAST search gave as top hits other glucokinases as top hitswith e-values from 3E-170 to 1E-30 and around 35% to 65% identity (NCBI: B2J224). A PDB BLAST search gave as the top hit an article on glucokinase with an e-value of 9.33051E-39 and 32% identity. [[14]](http://198.128.26.117/genome/converter.php#ref-14) Out of the active sites mentioned, 8 out of 10 matched. The mismatches were V141N and G156S. The article did not mention any specific function of these residues, so a multiple sequence alignment was done on T-Coffee against reviewed cyanobacterial FASTA sequences from UniProt. This showed that at site 141, V141 was not found among the other cyanobacteria, nor N141. The other cyanobacteria showed the Q141 residue to be conserved. At site 156, the G156 was changed to S156, and the S156 residue was completely conserved among other cyanobacteria. A reverse BLAST search also gave the gene as a top hit, with an e-value of 0. From these observations, it was concluded that the gene most likely codes for glucokinase.

**Pyruvate Kinase**

Gene OID: 2505167661 EC: 2.7.1.40

Gene is predicted to code for pyruvate kinase. A NCBI BLAST returned a near-perfect match with another pyruvate kinase, and gave as top results other pyruvate kinases. The top hit had an e-value of 0 and a query coverage of 98% [[9]](http://198.128.26.117/genome/converter.php#ref-9). A reverse BLAST search also gave the gene as a top hit, with an e-value of 0. A PDB BLAST search gave an article on pyruvate kinase structure in E. coli. This particular hit had an e-value of 9.39729E-98 and a query coverage of 42%. All the active sites mentioned in this article matched with the active sites from *Microcoleus vaginatus*[[15]](http://198.128.26.117/genome/converter.php#ref-15).

**3-phosphoglycerate kinase**

Gene OID: 2505166234 EC 2.7.2.3

This gene has been predicted to encode for phosphoglycerate kinase, the enzyme responsible for function 2.7.2.3. An NCBI BLAST search gave other phosphoglycerate kinase enzymes as top hits, with an e-value of 0 and a query coverage of 100% (NCBI: Q3MF40). A reverse BLAST search also gave the gene as a top hit, with an e-value of 0. A PDB BLAST Search gave other phosphoglycerate kinase enzymes as top hits, with an e-value of 4.39465E-113 and a query coverage of 53%. [[2]](http://198.128.26.117/genome/converter.php#ref-2) Almost all active sites matched, except for one, with D199 being replaced with S199, which only differ in an oxygen group on the side chain.

**Xylulose-5-Phosphate Phosphoketolase**

Gene OID: 2505165738 EC: 4.1.2.9

This gene was predicted to code for phosphoketolase. A NCBI BLAST search gave as top hits phosphoketolaseswith e-values of 8e-22 to 8e-08 and around 22% identity. [[19]](http://198.128.26.117/genome/converter.php#ref-19) A PDB BLAST search gave as the top hits fructose-6-phosphate phosphoketolases. This was because PDB has no articles on EC 4.1.2.9, xylulose-5-phosphate phosphoketolase, in its database. Therefore, from the hits with the NCBI SwissProt database, it was predicted that this gene codes for phosphoketolase. As a final check, a reverse BLAST search was done, and this gene was the second hit. The top hit was gene 2505165968. This gene was checked as well to see if it coded for X5P phosphoketolase.

Gene OID: 2505165968 EC: 4.1.2.9

This gene was also predicted to code for phosphoketolase. A NCBI BLAST search gave as top hits probable phosphoketolases, and it also gave a X5P phosphoketolase as a top hit, with an e-value of 0 and 51% identity. [[16]](http://198.128.26.117/genome/converter.php#ref-16)

**Phosphopyruvate hydratase**

Gene OID: 2505166395 EC: 4.2.1.11

The gene is predicted to code for the Phosphopyruvate hydratase enzyme (EC 4.2.1.11). A NCBI BLAST (NCBI: Q110V4) search against the Swissprot Database gives other Phosphopyruvate hydratase enzymes top hits, with an e-value of 0 and a query coverage of 99%. A PDB BLAST [[11]](http://198.128.26.117/genome/converter.php#ref-11) search gave an article on the Phosphopyruvate hydratase enzyme from *E. coli*, and eight out of ten of the active sites mentioned in this article matched with the ones in *Microcoleus vaginatus*. This particular PDB hit had an e-value of 1.2769E-131 and an identity of 60%. A reverse BLAST search also gave the gene as a top hit, with an e-value of 0.

**Ribulose-phosphate 3-epimerase**

Gene OID: 2505170722 EC: 5.1.3.1

Gene was predicted to code for ribulose-phosphate 3-epimerase. A NCBI BLAST [[10]](http://198.128.26.117/genome/converter.php#ref-10) search gave as top results other ribulose-phosphate 3-epimerases, and the matches were very good, with an e-value of 1e-142 and a query coverage of 98%. A PDB BLAST [[27]](http://198.128.26.117/genome/converter.php#ref-27) search gave an article on the ribulose-phosphate 3-epimerase structure in *Synechocystis sp.*, a cyanobacteria. The alignment showed that the match was very good, with an e-value of 3.30682E-112 and a percent identity of 84%. Additionally, all the checked active sites matched up with each other. A reverse BLAST search also gave the gene as a top hit, with an e-value of 0.

**Phosphoglycerate mutase**

Gene OID 2505170019, 2505169225, 2505165971, and 2505168909 EC 5.4.2.1

Gene OID 2505170019

This gene is predicted to code for a protein associated with phosphogylcerate mutase behavior, a claim supported by close homologs with other 5.4.2.1 enzymes found in PDB [[25]](http://198.128.26.117/genome/converter.php#ref-25) and Swissprot databases (NCBI: A4W6B3). However, because the protein has twice as many amino acids as 5.4.2.1 enzymes and multiple alignment via T-coffee leaves a wide region without any matches. (NCBI hit had a e-value of 5e-18 and a query coverage of 89%)

Literature also suggests that distinguishing phosphoglycerate mutase from other homologous enzymes can be problematic, so further research may be necessary to completely understand this gene/process. A reverse BLAST search gave the gene as a top hit, with an e-value of 3e-18.

Gene OID: 2505169225

The mis-annotated gene was found via an IMG Genome BLAST against 2505170019. It was confirmed to code for phosphoglycerate mutase, with reasonably good active site matches

on PDB BLAST [[24]](http://198.128.26.117/genome/converter.php#ref-24). An NCBI BLAST with swissprot [[6]](http://198.128.26.117/genome/converter.php#ref-6) revealed other phosphoglycerate mutases in other organisms, all of the biphosphoglycerate-dependent variety.

Gene OID: 2505165971

Gene 2505165971 was found to code for a cofactor dependent phosphoglycerate mutase. A NCBI BLAST was done, and the top results were a mixture of alpha-ribazole phosphatases and [probable] phosphoglycerate mutases. A PDB BLAST search [[17]](http://198.128.26.117/genome/converter.php#ref-17) gave an article about phosphatase PhoE (a member of the cofactor dependent phosphoglycerate superfamily) and the active site residues from this article were compared with the ones in *Microcoleus vaginatus.* Sites 16, 22 and 152 were mismatched. For site 16, N was replaced with S. The article says that N stabilizes the substrate due to its neutral side chain, and S has a neutral side chain as well. Furthermore, N and S are very similar in structure. A multiple sequence alignment was also done on T-Coffee, using FASTA sequences from UniProt, and this alignment showed that the S residue is completely conserved among the compared cyanobacteria sequences. For site 22, according to the article, the presence of a Q established enzyme identity as a PhoE, while the presence of T, S or C established identity as a cofactor dependent phosphoglycerate mutase. A multiple sequence alignment also showed the residue for cyanobacteria was conserved mostly as C, and as S in a few other cases, which is consistent with the article’s findings. Finally, for site 152, the G residue is not mentioned as having a specific function, and a multiple sequence alignment showed that the residue was completely conserved as K for cyanobacteria.

Gene OID 2505168909

This gene has been confirmed to encode for 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, the enzyme crucial for reaction 5.4.2.1. This conclusion is supported by very strong evidence for homology with other BPG-independent PGMs in NCBI BLAST and PDB BLAST [[5]](http://198.128.26.117/genome/converter.php#ref-5). In addition, active sites were a perfect match with the PDB sequence.

**Ribose 5-Phosphate Isomerase**

Gene OID: 2505165655 EC: 5.3.1.6

This enzyme and the gene coding for it were already found in the Calvin Cycle annotation. This entry is a copy of the description from the Calvin Cycle Annotation Report.

BLAST searches within IMG/ER and UniProt both gave ribose 5-phosphate isomerases as the top hits, as did NCBI BLAST (NCBI: Q111U2). returned an extremely good match with transketolase. A PDB BLAST [[4]](http://198.128.26.117/genome/converter.php#ref-4) search gave as a 1st result the article "A hyperthermostable D-ribose-5-phosphate isomerase from Pyrococcus horikoshii characterization and three-dimensional structure. (2002) Structure 10: 877-886".

The enzyme number matched and the organism used was Pyrococcus horikoshii.

The active sites were checked, and they all matched except for sites 100, 106 and 168.

In the organism, the sites were R100, M106 and D168. In microcoleus vaginatus, the sites were G100, R106 and D168. In the article, there was no mention of the functions of sites 106 or 168. It was said that site 100 was not completely conserved. Also, a multiple sequence alignment was showed in the article. This alignment showed that the aforementioned sites were not completely conserved, and the amino acids found in microcoleus vaginatus were found in some of the other organisms used for the alignment. To double-check, a multiple sequence alignment was also done against other similar genes from cyanobacteria. Site 100 was almost completely conserved along cyanobacteria as being a G, except for one cyanobacterium that was a T. Site 106 had many R's among other cyanobacteria, as well as Q and a single A. Site 168 had mostly A's among cyanobacteria, a couple D's, and one T. The few outliers could be explained as not actually being the correct enzyme, because some cyanobacteria listed multiple genes for enzyme number 5.3.1.6, which could be a mistake due to automatic annotation.

**Functional Studies**

There were many genes in this annotation report which were already talked about in other annotation reports, namely the Calvin cycle, pentose phosphate, heterolactic acid fermentation and glycolysis pathways. The only new enzymes are EC 2.7.2.1 and 4.1.2.22, which were both strong predicted to have genes coding for them.

**Discussion**

*Microcoleus vaginatus* possibly has all the enzymes needed for the bifidobacterium shunt. The most questionable enzyme is EC 1.1.1.27. Besides that, the other enzymes were predicted to have genes coding for them. Even if EC 1.1.1.27 doesn’t exist in *Microcoleus vaginatus*, the organism could still potentially use EC 1.1.1.28 to convert pyruvate to D-lactate instead of L-lactate. Experiments need to be done to see if the organism actually undergoes this pathway, because it is only known that *Microcoleus vaginatus* can undergo the bifidobacterium shunt, but not if it actually does. If *Microcoleus vaginatus* undergoes this fermentation pathway, it should be further studied if only this fermentation pathway is used, or if other fermentation pathways are used as well, and why that would be the case. Finally, studies need to be done to see if EC 1.1.1.27 exists in the organism and if it can create L-lactate, or if it can only create D-lactate with EC 1.1.1.28.

**References**

[1] Arai K, Ishimitsu T, Fushinobu S, Uchikoba H, Matsuzawa H, Taguchi H. "Active and inactive

state structures of unliganded Lactobacillus casei allosteric L-lactate dehydrogenase."

Proteins. Feb 15, 2010; 78(3):681-94.

[2] Auerbach G, Huber R, Grättinger M, Zaiss K, Schurig H, Jaenicke R, Jacob U. "Closed

structure of phosphoglycerate kinase from Thermotoga maritima reveals the catalytic

mechanism and determinants of thermal stability." Structure. Nov 15, 1997;

5(11):1475-83.

[3] Buss KA, Cooper DR, Ingram-Smith C, Ferry JG, Sanders DA, Hasson MS. "Urkinase:

structure of acetate kinase, a member of the ASKHA superfamily of

phosphotransferases." J Bacteriol. 2001 Jan; 183(2):680-6.

[4] Ishikawa K, Matsui I, Payan F, Cambillau C, Ishida H, Kawarabayasi Y, Kikuchi H, Roussel A.

"A hyperthermostable D-ribose-5-phosphate isomerase from Pyrococcus horikoshii

characterization and three-dimensional structure." Structure. 2002 Jun; 10(6):877-86.

[5] Jedrzejas MJ, Chander M, Setlow P, Krishnasamy G. "Mechanism of catalysis of the

cofactor-independent phosphoglycerate mutase from Bacillus stearothermophilus.

Crystal structure of the complex with 2-phosphoglycerate." J Biol Chem. Jul 28, 2000;

275(30):23146-53.

[6] Kaneko T, Nakamura Y, Sato S, Asamizu E, Kato T, Sasamoto S, Watanabe A, Idesawa K,

Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kiyokawa C, Kohara M, Matsumoto M,

Matsuno A, Mochizuki Y, Nakayama S, Nakazaki N, Shimpo S, Sugimoto M, Takeuchi C,

Yamada M, Tabata S. "Complete genome structure of the nitrogen-fixing symbiotic

bacterium Mesorhizobium loti." DNA Res. Dec 31, 2000; 7(6):331-8.

[7] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[8] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[9] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo

K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M,

Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[10] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo

K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M,

Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[11] Kühnel K, Luisi BF. "Crystal structure of the Escherichia coli RNA degradosome component

enolase." J Mol Biol. Oct 26, 2001; 313(3):583-92.

[12] Louis A, Ollivier E, Aude JC, Risler JL. "Massive sequence comparisons as a help in

annotating genomic sequences." Genome Res. 2001 Jul; 11(7):1296-303.

[13] Lucas J. Stal, Roy Moezelaar"Fermentation in cyanobacteria" FEMS Microbiology Reviews.

0; 21(2):179-211.

[14] Lunin VV, Li Y, Schrag JD, Iannuzzi P, Cygler M, Matte A. "Crystal structures of Escherichia

coli ATP-dependent glucokinase and its complex with glucose." J Bacteriol. 2004 Oct;

186(20):6915-27.

[15] Mattevi A, Valentini G, Rizzi M, Speranza ML, Bolognesi M, Coda A. "Crystal structure of

Escherichia coli pyruvate kinase type I: molecular basis of the allosteric transition."

Structure. Jul 15, 1995; 3(7):729-41.

[16] Nakamura Y, Kaneko T, Sato S, Mimuro M, Miyashita H, Tsuchiya T, Sasamoto S,

Watanabe A, Kawashima K, Kishida Y, Kiyokawa C, Kohara M, Matsumoto M, Matsuno

A, Nakazaki N, Shimpo S, Takeuchi C, Yamada M, Tabata S. "Complete genome

structure of Gloeobacter violaceus PCC 7421, a cyanobacterium that lacks thylakoids."

DNA Res. Aug 31, 2003; 10(4):137-45.

[17] Rigden DJ, Littlejohn JE, Henderson K, Jedrzejas MJ. "Structures of phosphate and

trivanadate complexes of Bacillus stearothermophilus phosphatase PhoE: structural and

functional analysis in the cofactor-dependent phosphoglycerate mutase superfamily." J

Mol Biol. Jan 17, 2003; 325(3):411-20.

[18] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[19] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[20] Schörken U, Thorell S, Schürmann M, Jia J, Sprenger GA, Schneider G. "Identification of

catalytically important residues in the active site of Escherichia coli transaldolase." Eur J

Biochem. 2001 Apr; 268(8):2408-15.

[21] Takahashi K, Tagami U, Shimba N, Kashiwagi T, Ishikawa K, Suzuki E. "Crystal structure of

Bifidobacterium Longum phosphoketolase; key enzyme for glucose metabolism in

Bifidobacterium." FEBS Lett. Sep 24, 2010; 584(18):3855-61.

[22] Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, Masui N, Fuji F, Hirama C, Nakamura

Y, Ogasawara N, Kuhara S, Horikoshi K. "Complete genome sequence of the alkaliphilic

bacterium Bacillus halodurans and genomic sequence comparison with Bacillus subtilis."

Nucleic Acids Res. Nov 1, 2000; 28(21):4317-31.

[23] Thorell S, Schürmann M, Sprenger GA, Schneider G. "Crystal structure of decameric

fructose-6-phosphate aldolase from Escherichia coli reveals inter-subunit helix swapping

as a structural basis for assembly differences in the transaldolase family." J Mol Biol. May

24, 2002; 319(1):161-71.

[24] Uhrínová S, Uhrín D, Nairn J, Price NC, Fothergill-Gilmore LA, Barlow PN. "Solution structure

and dynamics of an open beta-sheet, glycolytic enzyme, monomeric 23.7 kDa

phosphoglycerate mutase from Schizosaccharomyces pombe." J Mol Biol. Feb 16, 2001;

306(2):275-90.

[25] Watkins HA, Baker EN. "Structural and functional analysis of Rv3214 from Mycobacterium

tuberculosis, a protein with conflicting functional annotations, leads to its characterization

as a phosphatase." J Bacteriol. 2006 May; 188(10):3589-99.

[26] Wikner C, Nilsson U, Meshalkina L, Udekwu C, Lindqvist Y, Schneider G. "Identification of

catalytically important residues in yeast transketolase." Biochemistry. Dec 16, 1997;

36(50):15643-9.

[27] Wise EL, Akana J, Gerlt JA, Rayment I. "Structure of D-ribulose 5-phosphate 3-epimerase

from Synechocystis to 1.6 A resolution." Acta Crystallogr D Biol Crystallogr. 2004 Sep;

60(Pt 9):1687-90.

## Heterolactic Acid Fermentation

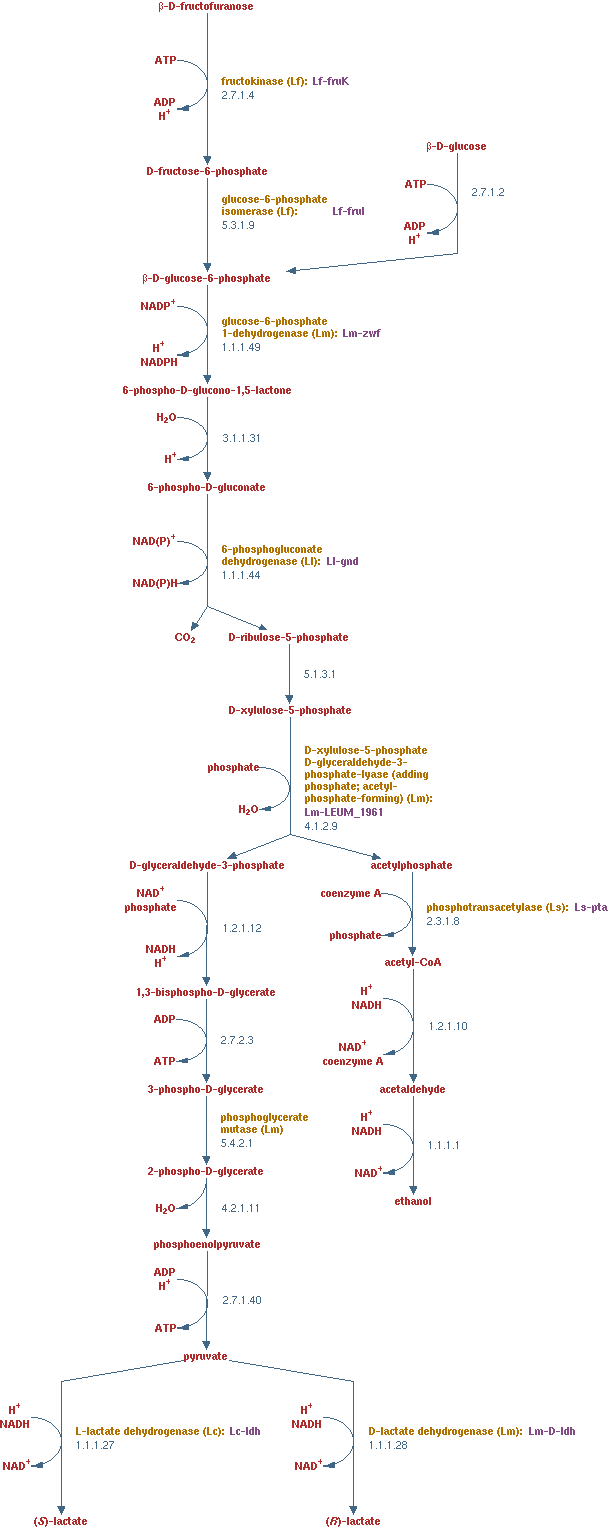
Many different fermentation pathways are well-documented in cyanobacteria

[[24]](http://198.128.27.8/genome/converter.php#ref-24). One of these well-documented fermentation pathways is heterolactic fermentation.

In most heterofermentative strains sugars are transported into the cell by a non-phosphorylating sugar permease system. After entering the cell the intracellullar sugars are phosphorylated by a dedicated kinase, such as glucokinase or fruktokinase. In the case of glucose , the phosphorylation results in the formation of β-D-glucose-6-phosphate . The phosphorylated forms of other sugars (such as D-fructose-6-phosphate ) are also converted to β-D-glucose-6-phosphate , which is channeled into the phosphoketolase pathway, resulting in the formation of D-glyceraldehyde-3-phosphate , acetylphosphate , CO2 and two equivalents of reducing power, in the form of NAD(P)H. These two equivalents of reducing power are then reoxidized via acetyl-CoA (derived from the acetyl-phosphate) yielding ethanol . D-glyceraldehyde-3-phosphate is processed by glycolytic enzymes to pyruvate and then lactate , generating two molecules of ATP. Since one ATP was used in the initial phosphorylation of the sugar, the net gain is only one ATP per sugar.

A list of the enzymes necessary for this pathway was made with the help of MetaCyc. From this list, two enzymes were definitely missing, such that *Microcoleus vaginatus* probably does not undergo heterolactic acid fermentation. All of the other enzymes necessary for the pathway were found though.

**Automated Annotation**



S Fig. 1: MetaCyc pathway map of heterolactic acid fermentation. Original image from http://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=P122-PWY&detail-level=2&detail-level=1

**Ribulose-phosphate 3-epimerase**

Gene OID: 2505170722 EC: 5.1.3.1

Gene was confirmed to code for ribulose-phosphate 3-epimerase. A NCBI BLAST [[16]](http://198.128.27.8/genome/converter.php#ref-16) search gave as top results other ribulose-phosphate 3-epimerases, and the matches were very good according to the e values and query coverage. A PDB BLAST [[37]](http://198.128.27.8/genome/converter.php#ref-37) search gave an article on the ribulose-phosphate 3-epimerase structure in *Synechocystis sp.*, a cyanobacteria. The alignment showed that the match was very good, and additionally all the checked active sites matched up with each other.

**6-Phosphogluconate Dehydrogenase, Decarboxylating**

Gene OID: 2505169298 EC: 1.1.1.44

Gene was confirmed to code for 6-phosphogluconate dehydrogenase. A NCBI BLAST [[19]](http://198.128.27.8/genome/converter.php#ref-19) search was done and the top hits were other 6-phosphogluconate dehydrogenases. The e-values for these hits were very low and the query coverage was high. Additionally, a PDB BLAST [[7]](http://198.128.27.8/genome/converter.php#ref-7) search was done and an article on the 6-phosphogluconate dehydrogenase structure in geobaccilus stearothermophilus was found. The active sites in this article were compared to the active site residues in *Microcoleus vaginatus*, and all of them matched.

**6-Phosphogluconolactonase**

Gene OID: 2505170610 EC: 3.1.1.31

Gene was confirmed to code for 6-phosphogluconolactonase. A NCBI BLAST

[[14]](http://198.128.27.8/genome/converter.php#ref-14) search gave as top hits other 6-phosphogluconolactonases with a high degree of homology. A PDB BLAST [[10]](http://198.128.27.8/genome/converter.php#ref-10) search gave an article on the 6-phosphogluconolactonase structure from *Trypanosoma brucei.* A check of the active site residues showed that all active site residues matched between the ones described in the article and the ones in *Microcoleus vaginatus*.

**Glucose-6-phosphate 1-dehydrogenase**

Gene OID: 2505169095 EC: 1.1.1.49

Gene is confirmed to code for glucose-6-phosphate 1-dehydrogenase, with almost no ambiguity. NCBI BLAST [[17]](http://198.128.27.8/genome/converter.php#ref-17) reveals near-perfect matches with other G6PD enzymes, including from the cyanobacterium Synechosystis, and PDB BLAST [[9]](http://198.128.27.8/genome/converter.php#ref-9) reveals a closely related G6PD in *L. Mesenteroides* with a perfect active site match.

**Pyruvate Kinase**

Gene OID: 2505167661 EC: 2.7.1.40

Gene is confirmed to code for pyruvate kinase. A NCBI BLAST returned a near-perfect match with another pyruvate kinase [[15]](http://198.128.27.8/genome/converter.php#ref-15), and gave as top results other pyruvate kinases. A PDB BLAST [[27]](http://198.128.27.8/genome/converter.php#ref-27) search gave an article on pyruvate kinase structure in E. coli. All the active sites mentioned in this article matched with the active sites from *Microcoleus vaginatus.*

**Phosphopyruvate hydratase**

Gene OID: 2505166395 EC: 4.2.1.11

The gene (2505166395) is confirmed to code for the Phosphopyruvate hydratase enzyme (EC 4.2.1.11). A NCBI BLAST (NCBI: Q110V4) search against the swissprot database gives other Phosphopyruvate hydratase enzymes top hits. A PDB BLAST [[20]](http://198.128.27.8/genome/converter.php#ref-20) search gave an article on the Phosphopyruvate hydratase enzyme from *E. coli*, and eight out of ten of the active sites mentioned in this article matched with the ones in *Microcoleus vaginatus*.

**Phosphoglycerate mutase**

Gene OID 2505170019, 2505169225, 2505165971, and 2505168909 EC 5.4.2.1

Gene OID 2505170019

This gene definitely encodes a protein associated with phosphogylcerate mutase behavior, a claim supported by close homologs with other 5.4.2.1 enzymes found in PDB

[[35]](http://198.128.27.8/genome/converter.php#ref-35) and Swissprot databases (NCBI: A4W6B3). However, because the protein has twice as many amino acids as 5.4.2.1 enzymes and multiple alignment via T-coffee leaves a wide region without any matches.

Literature [[36]](http://198.128.27.8/genome/converter.php#ref-36) also suggests that distinguishing phosphoglycerate mutase from other homologous enzymes can be problematic, so further research may be necessary completely understanding this gene/process is crucial.

Gene OID: 2505169225

The misannotated gene was found via an IMG Genome BLAST against 2505170019. It was confirmed to code for phosphoglycerate mutase, with reasonably good active site matches on PDB BLAST [[34]](http://198.128.27.8/genome/converter.php#ref-34). An NCBI BLAST with the SwissProt database [[13]](http://198.128.27.8/genome/converter.php#ref-13) revealed other phosphoglycerate mutases in other organisms, all of the biphosphoglycerate-dependent variety.

Gene OID: 2505165971

Gene 2505165971 was found to code for a cofactor dependent phosphoglycerate mutase. A NCBI BLAST was done, and the top results were a mixture of alpha-ribazole phosphatases and [probable] phosphoglycerate mutases. A PDB BLAST search [[30]](http://198.128.27.8/genome/converter.php#ref-30) gave an article about phosphatase PhoE (a member of the cofactor dependent phosphoglycerate superfamily) and the active site residues from this article were compared with the ones in microcoleus vaginatus. Sites 16, 22 and 152 were mismatched. For site 16, N was replaced with S. The article says that N stabilizes the substrate due to its neutral side chain, and S has a neutral side chain as well. Furthermore, N and S are very similar in structure. A multiple sequence alignment was also done on T-Coffee, using FASTA sequences from UniProt, and this alignment showed that the S residue is completely conserved among the compared cyanobacteria sequences. For site 22, according to the article, the presence of a Q established enzyme identity as a PhoE, while the presence of T, S or C established identity as a cofactor dependent phosphoglycerate mutase. A multiple sequence alignment also showed the residue for cyanobacteria was conserved mostly as C, and as S in a few other cases, which is consistent with the article’s findings. Finally, for site 152, the G residue is not mentioned as having a specific function, and a multiple sequence alignment showed that the residue was completely conserved as K for cyanobacteria.

Gene OID 2505168909

This gene has been confirmed to encode for 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, the enzyme crucial for reaction 5.4.2.1. This conclusion is supported by very strong evidence for homology with other BPG-independent PGMs in NCBI BLAST and PDB BLAST [[11]](http://198.128.27.8/genome/converter.php#ref-11). In addition, active sites were a perfect match with the PDB sequence.

**Glyceraldehyde-3-Phosphate Dehydrogenase Type 1**

Gene OID: 2505170636 EC: 1.2.1.12

Gene is confirmed to code for glyceraldehyde-3-phosphate dehydrogenase (GAPD). A NCBI BLAST search gave as top results other GAPDs. A PDB BLAST [[38]](http://198.128.27.8/genome/converter.php#ref-38) search gave an article with active site residues for a GAPD, and NCBI BLAST provided very strong evidence for the same conclusion [[26]](http://198.128.27.8/genome/converter.php#ref-26). The active site residues all matched with the ones in *Microcoleus vaginatus*.

**3-phosphoglycerate kinase**

Gene OID 2505166234 EC 2.7.2.3

This gene has been confirmed to encode for phosphoglycerate kinase, the enzyme responsible for function 2.7.2.3. Strong homology between phosphoglycerate kinase in other annotated genomes in the Swissprot database (NCBI:Q3MF40) and PDB [[4]](http://198.128.27.8/genome/converter.php#ref-4), including close active site matching, confirms this. Only one active site does not match, with D199 being replaced with S199, which only differ in an oxygen group on the side chain.

**Fructokinase**

Gene OID: 2505166498 EC 2.7.1.4

Gene was predicted to code for fructokinase. A NCBI BLAST search [[12]](http://198.128.27.8/genome/converter.php#ref-12) gave as top hits other fructokinases, with e-values from 1E-50 to 1E-28 and around 35% identity. A PDB BLAST search gave as a top hit an article on fructokinase structure [[8]](http://198.128.27.8/genome/converter.php#ref-8), with an e-value of 1.68757E-26 and 28% identity. There were two binding sites mentioned in this article: one for ATP, and one for the substrate. For the ATP binding site, 6 out of 12 active sites matched. The mismatches were C183N, D215E, A267V, L297Y, N299S and F304M. These residues were not mentioned to have specific function, so a multiple sequence alignment was done on T-Coffee with cyanobacterial FASTA seqeuences from UNIPROT. This showed that for site 183, C183 was not found for other cyanobacteria, and the N183 residue was mostly conserved. At site 215, the E215 was completely conserved among other cyanobacteria. At site 267, A267 was not found for other cyanobacteria, and the V267 residue was highly conserved. At site 297, the L297 residue was not found among other cyanobacteria, but the Y297 was found for some other cyanobacteria. At site 299, the N299 residue was not found for any other cyanobacteria, but the S299 residue was found. For site F304, the alignment was not helpful, since no other cyanobacteria had the F or M residue at this spot.

For the substrate binding site, 6 out of 8 active sites matched. The mismatches, S56A and F153L, were also analyzed with the same multiple seqeunce alignment. At site 56, the S56 residue was only found for one other cyanobacteria, while the rest of the residues were A56. At site 153, the F153 residue was not found for any other cyanobacteria, but the L153 residue was found for a few of the cyanobacteria.

From the active site data, it seems that this protein is better at binding its substrate than ATP, which seems reasonable since the protein needs to be able to bind to its substrate more than to ATP. The multiple sequence alignment data showed the mismatches to be reasonable, such that it can still be predicted with a good degree of certainty that the gene codes for a fructokinase.

**Glucose-6-phosphate Isomerase**

Gene OID: 2505167736 EC: 5.3.1.9

Gene is confirmed to code for Glucose-6-phosphate Isomerase (GPI). A NCBI BLAST gives as top hits other GPIs, and the e-values and query coverage are both very good

(NCBI: B2J5F1). A PDB BLAST gave an article [[3]](http://198.128.27.8/genome/converter.php#ref-3) about the active site residues for GPI. There were some active site residues that didn’t match, namely T260, F261 and T262. In *Microcoleus vaginatus*, these active site residues were S260, G261 and G262. For site 260, T and S are similar in structure, and the article says that the T is important for its side chain hydroxyl group, which S also has. Also, a multiple sequence alignment was done with FASTA sequences from UniProt, and at site 260 the S residue is completely conserved when compared to other cyanobacteria. For sites 261 and 262, nothing was mentioned in the article, so these sites can be assumed to be less important for the active site conS Figuration. Also, a multiple sequence alignment done on T-Coffee showed that for a vast majority of cyanobacteria, the GG sequence was conserved.

**Glucokinase**

Gene OID: 2505166819 EC: 2.7.1.2

This gene may code for a glucokinase, but it is unclear. A NCBI BLAST search gave as top hits other glucokinases [[1]](http://198.128.27.8/genome/converter.php#ref-1) with e-values between 5E-43 and 5E-39 and around 35% identity, but all the hits after these top three hits were not for glucokinases. A PDB BLAST search gave as the top hit an unpublished article on glucokinase (PDB: 2QM1) with an e-value of 1.98516E-47 and 36%, but all the hits after that one were for other proteins besides glucokinases. Therefore, it is unclear if these gene actually codes for a glucokinase or not. Further research needs to be done.

Gene OID: 2505170243 EC: 2.7.1.2

This gene was predicted to code for glucokinase. A NCBI BLAST search gave as top hits other glucokinases as top hits [[33]](http://198.128.27.8/genome/converter.php#ref-33) with e-values from 3E-170 to 1E-30 and around 35% to 65% identity. A PDB BLAST search gave as the top hit an article on glucokinase

[[25]](http://198.128.27.8/genome/converter.php#ref-25) with an e-value of 9.33051E-39 and 32% identity. Out of the active sites mentioned, 8 out of 10 matched. The mismatches were V141N and G156S. The article did not mention any specific function of these residues, so a multiple sequence alignment was done on T-Coffee against reviewed cyanobacterial FASTA sequences from UniProt. This showed that at site 141, V141 was not found among the other cyanobacteria, nor N141. The other cyanobacteria showed the Q141 residue to be conserved. At site 156, the G156 was changed to S156, and the S156 residue was completely conserved among other cyanobacteria. From these observations, it was concluded that the gene most likely coded for glucokinase.

**Xylulose-5-Phosphate Phosphoketolase**

Gene OID: 2505165738 EC: 4.1.2.9

This gene was predicted to code for phosphoketolase. A NCBI BLAST search gave as top hits phosphoketolases [[32]](http://198.128.27.8/genome/converter.php#ref-32) with e-values of 8E-22 to 8E-08 and around 22% identity. A PDB BLAST search gave as the top hits fructose-6-phosphate phosphoketolases. This was because PDB has no articles on EC 4.1.2.9, xylulose-5-phosphate phosphoketolase, in its database. Therefore, from the hits with the NCBI SwissProt database, it was predicted that this gene codes for phosphoketolase. As a final check, a reverse BLAST search was done, and this gene was the second hit. The top hit was gene 2505165968. This gene was checked as well to see if it coded for X5P phosphoketolase

Gene OID: 2505165968 EC: 4.1.2.9

This gene was also predicted to code for phosphoketolase. A NCBI BLAST search gave as top hits probable phosphoketolases, and it also gave a X5P phosphoketolase as a top hit [[28]](http://198.128.27.8/genome/converter.php#ref-28), with an e-value of 0 and 51% identity.

**L-Lactate Dehdyrogenase**

Gene OID: 2505168756 EC: 1.1.1.27/1.1.1.32

This gene could code for L-lactate dehydrogenase. This gene was previously predicted to code for malate dehydrogenase, but there is a possibility it could code for a bifunctional gene. A NCBI BLAST search gave as top hits malate dehydrogenases, but there is one hit for L-lactate dehydrogenase [[23]](http://198.128.27.8/genome/converter.php#ref-23) with an e-value of 6E-96 and 47% identity. A PDB BLAST search also mainly gave hits to articles on malate dehydrogenase, but some hits were articles for L-lactate dehydrogenase, like a hit for an article on L-lactate dehdyrogenase in *Lactobacillus casei* [[2]](http://198.128.27.8/genome/converter.php#ref-2) with an e-value of 3.05424E-41 and 29% identity. A reverse BLAST hit using the FASTA sequence from this hit gave gene 2505168756 as the top and only hit. Therefore, it is concluded that gene 2505168756 might code for a protein bifunctional for EC 1.1.1.27 and 1.1.1.32 function.

**D-Lactate Dehydrogenase**

Gene OID: 2505170579 EC: 1.1.1.28

This gene was confirmed to code for D-lactate dehydrogenase. A NCBI BLAST search gave D-lactate dehydrogenases as top hits [[6]](http://198.128.27.8/genome/converter.php#ref-6) with e-values from 1E-126 to 1E-44 and around 40% to 50% identity. A PDB BLAST search gave an article on D-lactate dehdyrogenase as a top hit [[29]](http://198.128.27.8/genome/converter.php#ref-29) with an e-value of 9.60815E-49 and 35% identity. The article mentioned active sites, and 8 out 11 of these matched. The mismatches were V233T, H155K and N212T. The article mentioned V233 being important to to its hydrophobicity, but T233 would not be hydrophobic. The article also mentioned H155 being important for its imidazole ring to interact with bridging water molecules. K155 doesn’t have an imidazole ring, but it also has a side chain nitrogen that could similar interact with bridging water molecules. N212 was mentioned to be important for participating in 2 hydrogen bonds, which T212 could do as well, since part of its side chain (the NH2 group) can participate in hydrogen bonding. As further confirmation, a multiple sequence alignment was done on T-Coffee with cyanobacterial FASTA sequences from UniProt. This showed that at site 155, no other cyanobacteria had H155, and all but one of them had the K155 residue. At site 212, the N212 residue was not found among any other cyanobacteria, while all but one cyanobacteria had the T212 residue. At site V233, no other cyanobacteria had residue V233, while all but one had residue T233. All these considerations taken together leads to the conclusion that the gene probably codes for D-lactate dehydrogenase, along with the fact that a reverse BLAST gave the same gene as the top hit.

**Alcohol dehydrogenase**

Gene OID: 2505166244 EC: 1.1.1.1

Gene is predicted to code for alcohol dehydrogenase. A NCBI BLAST search gave other alcohol dehydrogenases as top hits [[5]](http://198.128.27.8/genome/converter.php#ref-5) with e-values from 0 to 2E-100 and around 60% to 70% identity. A PDB BLAST search gave an article on alcohol dehydrogenase [[31]](http://198.128.27.8/genome/converter.php#ref-31) with an e-value of 4.76389E-137 and 63% identity. All the active sites mentioned in the article matched, and a reverse BLAST search gave gene 2505166244 as the top hit with an e-value of 0. Therefore, the gene is strongly predicted to code for alcohol dehydrogenase.

**Acetaldehyde dehydrogenase**

Gene OID: 2505170539 EC: 1.2.1.10

No gene in *Microcoleus vaginatus* was predicted by IMG/ER to code for aldehyde dehydrogenase. A search was done on UniProt for cyanobacterial FASTA sequences of aldehyde dehydrogenases, and these sequences were BLASTED against the *Microcoleus vaginatus* genome. The top hit was gene 2505170539, a gene predicted to code for succinate-semialdehyde dehydrogenase (EC 1.2.1.16). A NCBI BLAST search gave succinate-semialdehyde dehydrogenases as top hits [[18]](http://198.128.27.8/genome/converter.php#ref-18) with e-values ranging from 0 to 1E-77 and ranging from 33% to 64% identity. A PDB BLAST search gave as top hits articles on semialdehyde dehydrogenases [[22]](http://198.128.27.8/genome/converter.php#ref-22), with the best hit having an e-value of 1.00415E-70 and 35% identity. A quick look at the active sites showed good matching between the PDB and gene FASTA sequences. A reverse BLAST search gave gene 2505170539 as the top hit. This suggests that the gene actually codes for a protein with function for EC 1.2.1.16, and it seems that there is no gene in *Microcoleus vaginatus* that codes for EC 1.2.1.10.

**Phosphate acetyltransferas**

EC 2.3.1.8

IMG/ER predicted no genes within *Microcoleus vaginatus* to code for EC 2.3.1.8. This enzyme was then found on UNIPROT and the results were limited to the genes from only cyanobacteria coding for this enzyme. The FASTA sequence of a gene from *Synechocystis* sp. (strain ATCC 27184 / PCC 6803 / N-1) was BLAST-ed against the *Microcoleus vaginatus* PCC 9802 genome on IMG/ER. The top result from this BLAST search, with an e-value of 7e-18, was gene 2505166026.

Gene 2505166026 (listed as BioD-like N-terminal domain of phosphotransacetylase) has no EC number linked to it within IMG/ER. The NCBI BLAST search gave many different enzymes as hits, including uncharacterized proteins, phosphate acetyltransferase, Cobyric acid synthase, and ATP-dependent dethiobiotin synthetase BioD. The top NCBI hit, with an e-value of 3e-138 and a query coverage of 94%, was an uncharacterized protein. [[21]](http://198.128.27.8/genome/converter.php#ref-21) The PDB BLAST search provided cluttered and unreliable results. Most likely, gene 2505166026 does not code for EC 2.3.1.8.

**Functional Studies**

Many of the genes annotated in this pathway showed up in the glycolysis and pentose phosphate pathways as well. The functional studies sections of those two reports should be read because many of the same problematic genes show up again in this report.

There are some genes that are new in this pathway, and some of these genes need further functional studies. For example, gene 2505166819 needs further studies done on it. This gene was predicted to code for glucokinase (EC 2.7.1.2). However, BLAST searches do not give other glucokinases as hits, so it seems that the gene does not code for a glucokinase. This needs to be verified. There is another gene confirmed to code for glucokinase, so it would make sense if this gene doesn’t code for glucokinase, since there is another gene that takes care of coding for glucokinase.

Gene 2505168756 should also be further analyzed. This gene was predicted to code for malate dehydrogenase (EC 1.1.1.32), but it could also be L-lactate dehydrogenase (EC 1.1.1.27). Studies need to be done to see if the gene codes only for malate dehydrogenase, or if it codes for a bifunctional protein. An experiment should be done to see if *Microcoleus vaginatus* can produce L-lactate, because if it can, then it indicates the presence of L-lactate dehydrogenase, and this gene would be the gene that codes for it.

**Discussion**

Almost all the enzymes necessary for the heterolactic acid fermentation pathway were found. There were two enzymes without any genes found to code for them: EC 1.2.1.10 and 2.3.1.8. This means that *Microcoleus vaginatus* cannot convert acetylphosphate to ethanol, so it cannot undergo the rightmost part of the pathway as illustrated in S Figure 1. Also, *Microcoleus vaginatus* might be missing EC 1.1.1.27, L-lactate dehydrogenase. This is less problematic, since the organism has EC 1.1.1.28, D-lactate dehydrogenase, so it is possible that it only produces D-lactate through this pathway.

In conclusion, *Microcoleus vaginatus* is not predicted to be able to undergo heterolactic acid fermentation, because it is missing two enzymes in the pathway, such that it cannot convert acetylphosphate to ethanol.

However, *Microcoleus vaginatus* has an enzyme, acetate kinase (EC 2.7.2.1) that can convert acetylphosphate to acetate. There is a possibility that *Microcoleus vaginatus* could then undergo a modified pathway, where acetylphosphate is converted to acetate but everything else stays the same. This is plausible since acetate is one of the most prevalent products of cyanobacterial fermentation [[24]](http://198.128.27.8/genome/converter.php#ref-24). Converting acetylphosphate to acetate would also be plausible since when the reversible reaction goes this way, one ADP is converted to one ATP. As fermentation is occuring in order to result in a net gain of ATP, this step would help to achieve that goal. Regardless of if *Microcoleus vaginatus* does this or not though, the bottom line is that it cannot undergo heterlactic acid fermentation, but it might undergo a variation of it.

**References**

[1] Angell S, Schwarz E, Bibb MJ. "The glucose kinase gene of Streptomyces coelicolor A3(2):

its nucleotide sequence, transcriptional analysis and role in glucose repression." Mol

Microbiol. 1992 Oct; 6(19):2833-44.

[2] Arai K, Ishimitsu T, Fushinobu S, Uchikoba H, Matsuzawa H, Taguchi H. "Active and inactive

state structures of unliganded Lactobacillus casei allosteric L-lactate dehydrogenase."

Proteins. Feb 15, 2010; 78(3):681-94.

[3] Arsenieva D, Appavu BL, Mazock GH, Jeffery CJ. "Crystal structure of phosphoglucose

isomerase from Trypanosoma brucei complexed with glucose-6-phosphate at 1.6 A

resolution." Proteins. 2009 Jan; 74(1):72-80.

[4] Auerbach G, Huber R, Grättinger M, Zaiss K, Schurig H, Jaenicke R, Jacob U. "Closed

structure of phosphoglycerate kinase from Thermotoga maritima reveals the catalytic

mechanism and determinants of thermal stability." Structure. Nov 15, 1997;

5(11):1475-83.

[5] Barber RD, Rott MA, Donohue TJ. "Characterization of a glutathione-dependent formaldehyde

dehydrogenase from Rhodobacter sphaeroides." J Bacteriol. 1996 Mar; 178(5):1386-93.

[6] Bunch PK, Mat-Jan F, Lee N, Clark DP. "The ldhA gene encoding the fermentative lactate

dehydrogenase of Escherichia coli." Microbiology. 1997 Jan; 143 ( Pt 1)():187-95.

[7] Cameron S, Martini VP, Iulek J, Hunter WN. "Geobacillus stearothermophilus

6-phosphogluconate dehydrogenase complexed with 6-phosphogluconate." Acta

Crystallogr Sect F Struct Biol Cryst Commun. May 1, 2009; 65(Pt 5):450-4.

[8] Chua TK, Seetharaman J, Kasprzak JM, Ng C, Patel BK, Love C, Bujnicki JM, Sivaraman J.

"Crystal structure of a fructokinase homolog from Halothermothrix orenii." J Struct Biol.

2010 Sep; 171(3):397-401.

[9] Cosgrove MS, Gover S, Naylor CE, Vandeputte-Rutten L, Adams MJ, Levy HR. "An

examination of the role of asp-177 in the His-Asp catalytic dyad of Leuconostoc

mesenteroides glucose 6-phosphate dehydrogenase: X-ray structure and pH

dependence of kinetic parameters of the D177N mutant enzyme." Biochemistry. Dec 12,

2000; 39(49):15002-11.

[10] Delarue M, Duclert-Savatier N, Miclet E, Haouz A, Giganti D, Ouazzani J, Lopez P, Nilges M,

Stoven V. "Three dimensional structure and implications for the catalytic mechanism of

6-phosphogluconolactonase from Trypanosoma brucei." J Mol Biol. Feb 23, 2007;

366(3):868-81.

[11] Jedrzejas MJ, Chander M, Setlow P, Krishnasamy G. "Mechanism of catalysis of the

cofactor-independent phosphoglycerate mutase from Bacillus stearothermophilus.

Crystal structure of the complex with 2-phosphoglycerate." J Biol Chem. Jul 28, 2000;

275(30):23146-53.

[12] Jiang H, Dian W, Liu F, Wu P. "Isolation and characterization of two fructokinase cDNA

clones from rice." Phytochemistry. 2003 Jan; 62(1):47-52.

[13] Kaneko T, Nakamura Y, Sato S, Asamizu E, Kato T, Sasamoto S, Watanabe A, Idesawa K,

Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kiyokawa C, Kohara M, Matsumoto M,

Matsuno A, Mochizuki Y, Nakayama S, Nakazaki N, Shimpo S, Sugimoto M, Takeuchi C,

Yamada M, Tabata S. "Complete genome structure of the nitrogen-fixing symbiotic

bacterium Mesorhizobium loti." DNA Res. Dec 31, 2000; 7(6):331-8.

[14] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[15] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo

K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M,

Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[16] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo

K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M,

Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[17] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo

K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M,

Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[18] Kaneko T, Tanaka A, Sato S, Kotani H, Sazuka T, Miyajima N, Sugiura M, Tabata S.

"Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp.

strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92%

of the genome." DNA Res. Aug 31, 1995; 2(4):153-66, 191-8.

[19] Kaneko T, Tanaka A, Sato S, Kotani H, Sazuka T, Miyajima N, Sugiura M, Tabata S.

"Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp.

strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92%

of the genome." DNA Res. Aug 31, 1995; 2(4):153-66, 191-8.

[20] Kühnel K, Luisi BF. "Crystal structure of the Escherichia coli RNA degradosome component

enolase." J Mol Biol. Oct 26, 2001; 313(3):583-92.

[21] Langendorf CG, Key TL, Fenalti G, Kan WT, Buckle AM, Caradoc-Davies T, Tuck KL, Law

RH, Whisstock JC. "The X-ray crystal structure of Escherichia coli succinic

semialdehyde dehydrogenase; structural insights into NADP+/enzyme interactions."

PLoS One. Feb 18, 2010; 5(2):e9280.

[22] Langendorf CG, Key TL, Fenalti G, Kan WT, Buckle AM, Caradoc-Davies T, Tuck KL, Law

RH, Whisstock JC. "The X-ray crystal structure of Escherichia coli succinic

semialdehyde dehydrogenase; structural insights into NADP+/enzyme interactions."

PLoS One. Feb 18, 2010; 5(2):e9280.

[23] Louis A, Ollivier E, Aude JC, Risler JL. "Massive sequence comparisons as a help in

annotating genomic sequences." Genome Res. 2001 Jul; 11(7):1296-303.

[24] Lucas J. Stal, Roy Moezelaar"Fermentation in cyanobacteria" FEMS Microbiology Reviews.

0; 21(2):179-211.

[25] Lunin VV, Li Y, Schrag JD, Iannuzzi P, Cygler M, Matte A. "Crystal structures of Escherichia

coli ATP-dependent glucokinase and its complex with glucose." J Bacteriol. 2004 Oct;

186(20):6915-27.

[26] Martin W, Brinkmann H, Savonna C, Cerff R. "Evidence for a chimeric nature of nuclear

genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase

genes." Proc Natl Acad Sci U S A. Sep 15, 1993; 90(18):8692-6.

[27] Mattevi A, Valentini G, Rizzi M, Speranza ML, Bolognesi M, Coda A. "Crystal structure of

Escherichia coli pyruvate kinase type I: molecular basis of the allosteric transition."

Structure. Jul 15, 1995; 3(7):729-41.

[28] Posthuma CC, Bader R, Engelmann R, Postma PW, Hengstenberg W, Pouwels PH.

"Expression of the xylulose 5-phosphate phosphoketolase gene, xpkA, from Lactobacillus

pentosus MD363 is induced by sugars that are fermented via the phosphoketolase

pathway and is repressed by glucose mediated by CcpA and the mannose

phosphoenolpyruvate phosphotransferase system." Appl Environ Microbiol. 2002 Feb;

68(2):831-7.

[29] Razeto A, Kochhar S, Hottinger H, Dauter M, Wilson KS, Lamzin VS. "Domain closure,

substrate specificity and catalysis of D-lactate dehydrogenase from Lactobacillus

bulgaricus." J Mol Biol. Apr 19, 2002; 318(1):109-19.

[30] Rigden DJ, Littlejohn JE, Henderson K, Jedrzejas MJ. "Structures of phosphate and

trivanadate complexes of Bacillus stearothermophilus phosphatase PhoE: structural and

functional analysis in the cofactor-dependent phosphoglycerate mutase superfamily." J

Mol Biol. Jan 17, 2003; 325(3):411-20.

[31] Sanghani PC, Robinson H, Bosron WF, Hurley TD. "Human glutathione-dependent

formaldehyde dehydrogenase. Structures of apo, binary, and inhibitory ternary

complexes." Biochemistry. Sep 3, 2002; 41(35):10778-86.

[32] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[33] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[34] Uhrínová S, Uhrín D, Nairn J, Price NC, Fothergill-Gilmore LA, Barlow PN. "Solution structure

and dynamics of an open beta-sheet, glycolytic enzyme, monomeric 23.7 kDa

phosphoglycerate mutase from Schizosaccharomyces pombe." J Mol Biol. Feb 16, 2001;

306(2):275-90.

[35] Watkins HA, Baker EN. "Structural and functional analysis of Rv3214 from Mycobacterium

tuberculosis, a protein with conflicting functional annotations, leads to its characterization

as a phosphatase." J Bacteriol. 2006 May; 188(10):3589-99.

[36] Watkins HA, Baker EN. "Structural and functional analysis of Rv3214 from Mycobacterium

tuberculosis, a protein with conflicting functional annotations, leads to its characterization

as a phosphatase." J Bacteriol. 2006 May; 188(10):3589-99.

[37] Wise EL, Akana J, Gerlt JA, Rayment I. "Structure of D-ribulose 5-phosphate 3-epimerase

from Synechocystis to 1.6 A resolution." Acta Crystallogr D Biol Crystallogr. 2004 Sep;

60(Pt 9):1687-90.

[38] Yun M, Park CG, Kim JY, Park HW. "Structural analysis of glyceraldehyde 3-phosphate

dehydrogenase from Escherichia coli: direct evidence of substrate binding and

cofactor-induced conformational changes." Biochemistry. Sep 5, 2000; 39(35):10702-10.

## Homolactic Fermentation

Some cyanobacteria are predicted in the literature to undergo fermentation [[16]](http://198.128.27.8/genome/converter.php#ref-16). One of the most prevalent predicted fermentation pathways is homolactic fermentation, which has been observed in many cyanobacteria.

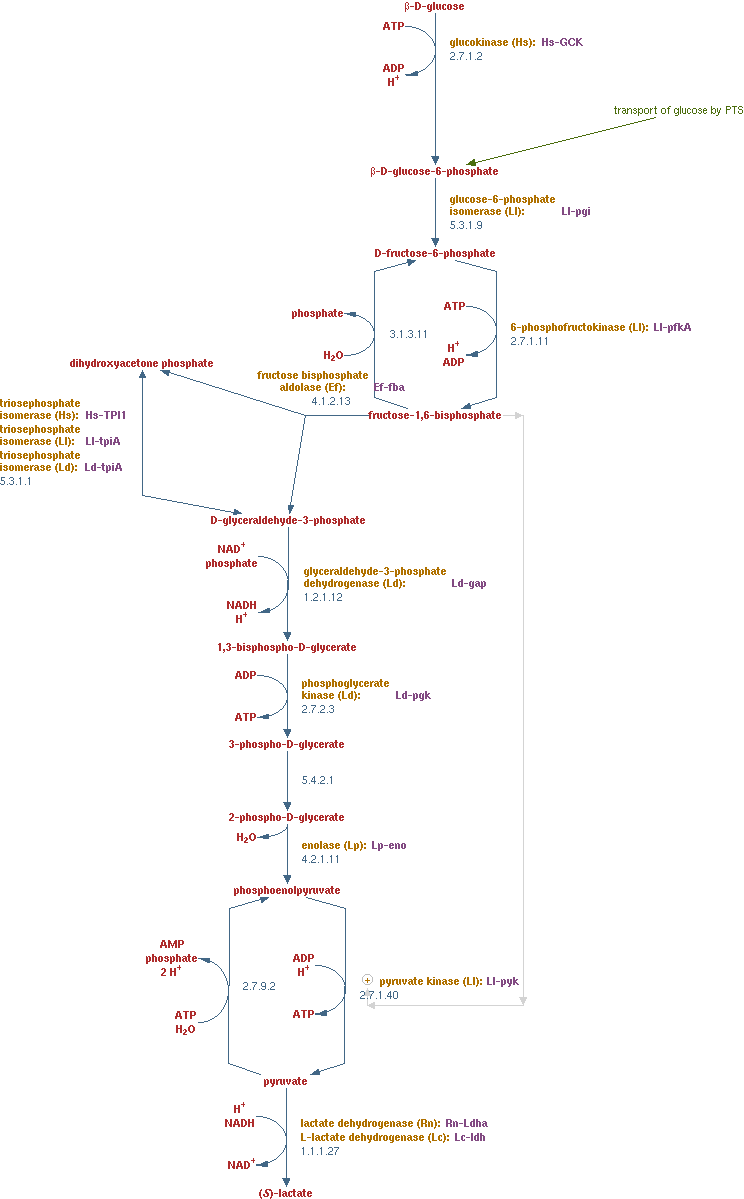
In homolactic fermentation sugars are metabolized via the Embden-Myerhof glycolytic pathway to pyruvate, which is then converted to lactate.

Sugars can enter bacterial cells in two different mechanisms. In most homofermentative strains sugars are transported by a sugar phosphotransferase system (PTS) that phosphorylates the sugar during transport. β-D-glucose is phosphorylated to β-D-glucose-6-phosphate , while D-fructose is usually phosphorylated to fructose-1-phosphate (although some PTSs were reported to generate D-fructose-6-phosphate. The second mechanism involves the transport of sugars via a permease, in which case they are not phosphorylated.

All sugars are then activated to fructose-1,6-bisphosphate through different routes, depending on the nature of the sugar and on the way it enters the cell. Intracellular fructose (that entered by a permease) is phosphorylated first to fructose-6-phosphate, then to fructose-1,6-biphosphate. Intracellular glucose (that entered by a permease) is first phosphorylated to glucose-6-phosphate, isomerized to fructose-6-phosphate, and finally phosphorylated to fructose-1,6-biphosphate.

Fructose-6-phosphate and fructose-1-phosphate (that entered via a PTS) are both phosphorylated to fructose-1,6-biphosphate by 6-phosphofructokinase and 1-phosphofructokinase , respectively, and glucose-6-phosphate (that entered via a PTS) is isomerized to fructose-6-phosphate, and finally phosphorylated to fructose-1,6-biphosphate.

The common product, fructose-1,6-biphosphate, is processed via the Embden-Meyerhof (EM) pathway to pyruvate, which is then converted to lactate, the the sole end-product of homolactic fermentation. Fructose-1,6-biphosphate is an important regulator of the pathway, being a positive effector of pyruvate kinase.

**Automated Annotation**

S Fig. 1: MetaCyc map of Homolactic Acid Fermentation Pathway. Full scale image at http://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=ANAEROFRUCAT-PWY&detail-level=2

**Glucokinase**

Gene OID: 2505166819 EC: 2.7.1.2

This gene may code for a glucokinase, but it is unclear. A NCBI BLAST search gave as top hits other glucokinases[[1]](http://198.128.27.8/genome/converter.php#ref-1) with e-values between 5E-43 and 5E-39 and around 35% identity, but all the hits after these top three hits were not for glucokinases. A PDB BLAST search gave as the top hit an unpublished article on glucokinase (PDB: 2QM1) with an e-value of 1.98516E-47 and 36%, but all the hits after that one were for other proteins besides glucokinases. Therefore, it is unclear if these gene actually codes for a glucokinase or not. Further research needs to be done.

Gene OID: 2505170243 EC: 2.7.1.2

This gene was predicted to code for glucokinase. A NCBI BLAST search gave as top hits other glucokinases as top hits[22] with e-values from 3E-170 to 1E-30 and around 35% to 65% identity. A PDB BLAST search gave as the top hit an article on glucokinase

[[17]](http://198.128.27.8/genome/converter.php#ref-17) with an e-value of 9.33051E-39 and 32% identity. Out of the active sites mentioned, 8 out of 10 matched. The mismatches were V141N and G156S. The article did not mention any specific function of these residues, so a multiple sequence alignment was done on T-Coffee against reviewed cyanobacterial FASTA sequences from UniProt. This showed that at site 141, V141 was not found among the other cyanobacteria, nor N141. The other cyanobacteria showed the Q141 residue to be conserved. At site 156, the G156 was changed to S156, and the S156 residue was completely conserved among other cyanobacteria. From these observations, it was concluded that the gene most likely coded for glucokinase.

**Glucose-6-phosphate Isomerase**

Gene OID: 2505167736 EC: 5.3.1.9

Gene is confirmed to code for Glucose-6-phosphate Isomerase (GPI). A NCBI BLAST gives as top hits other GPIs, and the e-values and query coverage are both very good

(NCBI: B2J5F1). A PDB BLAST gave an article [[3]](http://198.128.27.8/genome/converter.php#ref-3) about the active site residues for GPI. There were some active site residues that didn’t match, namely T260, F261 and T262. In *Microcoleus vaginatus*, these active site residues were S260, G261 and G262. For site 260, T and S are similar in structure, and the article says that the T is important for its side chain hydroxyl group, which S also has. Also, a multiple sequence alignment was done with FASTA sequences from UniProt, and at site 260 the S residue is completely conserved when compared to other cyanobacteria. For sites 261 and 262, nothing was mentioned in the article, so these sites can be assumed to be less important for the active site conS Figuration. Also, a multiple sequence alignment done on T-Coffee showed that for a vast majority of cyanobacteria, the GG sequence was conserved.

**L-Lactate Dehdyrogenase**

Gene OID: 2505168756 EC: 1.1.1.27/1.1.1.32

This gene could code for L-lactate dehydrogenase. This gene was previously predicted to code for malate dehydrogenase, but there is a possibility it could code for a bifunctional gene. A NCBI BLAST search gave as top hits malate dehydrogenases, but there is one hit for L-lactate dehydrogenase [[15]](http://198.128.27.8/genome/converter.php#ref-15) with an e-value of 6E-96 and 47% identity. A PDB BLAST search also mainly gave hits to articles on malate dehydrogenase, but some hits were articles for L-lactate dehydrogenase, like a hit for an article on L-lactate dehdyrogenase in *Lactobacillus casei* [[2]](http://198.128.27.8/genome/converter.php#ref-2) with an e-value of 3.05424E-41 and 29% identity. A reverse BLAST hit using the FASTA sequence from this hit gave gene 2505168756 as the top and only hit. Therefore, it is concluded that gene 2505168756 might code for a protein bifunctional for EC 1.1.1.27 and 1.1.1.32 function.

**Pyruvate Kinase**

Gene OID: 2505167661 EC: 2.7.1.40

Gene is confirmed to code for pyruvate kinase. A NCBI BLAST returned a near-perfect match with another pyruvate kinase [[11]](http://198.128.27.8/genome/converter.php#ref-11), and gave as top results other pyruvate kinases. A PDB BLAST [[19]](http://198.128.27.8/genome/converter.php#ref-19) search gave an article on pyruvate kinase structure in E. coli. All the active sites mentioned in this article matched with the active sites from *Microcoleus vaginatus.*

**Pyruvate, water dikinase**

Gene OID 2505165817, 2505167376, and 2505168646 EC: 2.7.9.2

Three genes (2505165817, 2505167376, and 2505168646) were predicted to code for the pyruvate, water dikinase enzyme (EC: 2.7.9.2). Based on numerical data from COG and Pfam, the enzyme 2.7.9.2 seems to be coded by either 2505165817 or 2505168646, but not 2505167376. PDB BLAST searches [[24]](http://198.128.27.8/genome/converter.php#ref-24) and NCBI BLASTs [[12]](http://198.128.27.8/genome/converter.php#ref-12) gave better numerical results for 2505168646 than 2505165817, but both of the active sites matched well for both of these genes from the PDB searches. Based on results from T-Coffee Multiple Sequence Alignment searches and an IMG Genome BLAST search, it is safe to say that genes 2505165817 and 2505168646 are very similar, but we can not be certain that they necessarily bear homology to one another. No clear conclusion can be made regarding gene 2505167376. Further research is needed.

**Phosphopyruvate hydratase**

Gene OID: 2505166395 EC: 4.2.1.11

The gene (2505166395) is confirmed to code for the Phosphopyruvate hydratase enzyme (EC 4.2.1.11). A NCBI BLAST (NCBI: Q110V4) search against the swissprot database gives other Phosphopyruvate hydratase enzymes top hits. A PDB BLAST [[13]](http://198.128.27.8/genome/converter.php#ref-13) search gave an article on the Phosphopyruvate hydratase enzyme from *E. coli*, and eight out of ten of the active sites mentioned in this article matched with the ones in *Microcoleus vaginatus*.

**Phosphoglycerate mutase**

Gene OID 2505170019, 2505169225, 2505165971, and 2505168909 EC 5.4.2.1

Gene OID 2505170019

This gene definitely encodes a protein associated with phosphogylcerate mutase behavior, a claim supported by close homologs with other 5.4.2.1 enzymes found in PDB

[[27]](http://198.128.27.8/genome/converter.php#ref-27) and Swissprot databases (NCBI: A4W6B3). However, because the protein has twice as many amino acids as 5.4.2.1 enzymes and multiple alignment via T-coffee leaves a wide region without any matches.

Literature [[26]](http://198.128.27.8/genome/converter.php#ref-26) also suggests that distinguishing phosphoglycerate mutase from other homologous enzymes can be problematic, so further research may be necessary completely understanding this gene/process is crucial.

Gene OID: 2505169225

The misannotated gene was found via an IMG Genome BLAST against 2505170019. It was confirmed to code for phosphoglycerate mutase, with reasonably good active site matches on PDB BLAST [[25]](http://198.128.27.8/genome/converter.php#ref-25). An NCBI BLAST with the SwissProt database [[8]](http://198.128.27.8/genome/converter.php#ref-8) revealed other phosphoglycerate mutases in other organisms, all of the biphosphoglycerate-dependent variety.

Gene OID: 2505165971

Gene 2505165971 was found to code for a cofactor dependent phosphoglycerate mutase. A NCBI BLAST was done, and the top results were a mixture of alpha-ribazole phosphatases and [probable] phosphoglycerate mutases. A PDB BLAST search

[[20]](http://198.128.27.8/genome/converter.php#ref-20) gave an article about phosphatase PhoE (a member of the cofactor dependent phosphoglycerate superfamily) and the active site residues from this article were compared with the ones in microcoleus vaginatus. Sites 16, 22 and 152 were mismatched. For site 16, N was replaced with S. The article says that N stabilizes the substrate due to its neutral side chain, and S has a neutral side chain as well. Furthermore, N and S are very similar in structure. A multiple sequence alignment was also done on T-Coffee, using FASTA sequences from UniProt, and this alignment showed that the S residue is completely conserved among the compared cyanobacteria sequences. For site 22, according to the article, the presence of a Q established enzyme identity as a PhoE, while the presence of T, S or C established identity as a cofactor dependent phosphoglycerate mutase. A multiple sequence alignment also showed the residue for cyanobacteria was conserved mostly as C, and as S in a few other cases, which is consistent with the article’s findings. Finally, for site 152, the G residue is not mentioned as having a specific function, and a multiple sequence alignment showed that the residue was completely conserved as K for cyanobacteria.

Gene OID 2505168909

This gene has been confirmed to encode for 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, the enzyme crucial for reaction 5.4.2.1. This conclusion is supported by very strong evidence for homology with other BPG-independent PGMs in NCBI BLAST and PDB BLAST [[7]](http://198.128.27.8/genome/converter.php#ref-7). In addition, active sites were a perfect match with the PDB sequence.

**3-phosphoglycerate kinase**

Gene OID 2505166234 EC 2.7.2.3

This gene has been confirmed to encode for phosphoglycerate kinase, the enzyme responsible for function 2.7.2.3. Strong homology between phosphoglycerate kinase in other annotated genomes in the Swissprot database (NCBI:Q3MF40) and PDB [[4]](http://198.128.27.8/genome/converter.php#ref-4), including close active site matching, confirms this. Only one active site does not match, with D199 being replaced with S199, which only differ in an oxygen group on the side chain.

**Glyceraldehyde-3-Phosphate Dehydrogenase Type 1**

Gene OID: 2505170636 EC: 1.2.1.12

Gene is confirmed to code for glyceraldehyde-3-phosphate dehydrogenase (GAPD). A NCBI BLAST search gave as top results other GAPDs. A PDB BLAST [[28]](http://198.128.27.8/genome/converter.php#ref-28) search gave an article with active site residues for a GAPD, and NCBI BLAST provided very strong evidence for the same conclusion [[18]](http://198.128.27.8/genome/converter.php#ref-18). The active site residues all matched with the ones in *Microcoleus vaginatus*.

**6-Phosphogluconolactonase**

Gene OID: 2505170610 EC: 3.1.1.31

Gene was confirmed to code for 6-phosphogluconolactonase. A NCBI BLAST

[[9]](http://198.128.27.8/genome/converter.php#ref-9) search gave as top hits other 6-phosphogluconolactonases with a high degree of homology. A PDB BLAST [[5]](http://198.128.27.8/genome/converter.php#ref-5) search gave an article on the 6-phosphogluconolactonase structure from *Trypanosoma brucei.* A check of the active site residues showed that all active site residues matched between the ones described in the article and the ones in *Microcoleus vaginatus*.

**6-Phosphofructokinase:**

Gene OID: 2505167140 EC: 2.7.1.11

Gene is confirmed to code for enzyme 2.7.1.11. A NCBI BLAST search [[10]](http://198.128.27.8/genome/converter.php#ref-10) gave as top results other 6-phosphofructokinases. A PDB BLAST search [[21]](http://198.128.27.8/genome/converter.php#ref-21) gave an article that described active sites for the enzyme. These active site residues were compared to the ones for the enzyme coded for by the gene in *Microcoleus vaginatus*. 12 out of the 15 active site residues matched. The ones that weren’t conserved were R72, T158 and E161. These residues were N72, A158 and S161 for *Microcoleus vaginatus*.

To double-check these active sites, a multiple sequence alignment was done with T-Coffee using FASTA sequences from UniProt for comparison. This showed that for site 72, many other cyanobacteria had N72 instead of R72. At site 158, all but one cyanobacteria had A158 instead of T158. For site 161, some of the other cyanobacteria had S161 instead of E161. All these substitutions were still considered good matches by T-Coffee (as indicated by color), and the 6-phosphofructokinase genes from different cyanobacteria matched up well.

**Fructose-Bisphosphate Aldolase**

Gene OID: 2505166282 and 2505169777 EC: 4.1.2.13

Two hits were found for this corresponding enzyme. Gene 2505166282 was predicted to be a type II Fructose-Bisphosphate Aldolase. Gene 2505169777 was also predicted to be a Fructose-Bisphosphate Aldolase, and perhaps an archaeal one.

For gene 2505166282, it was predicted to code for a type II Fructose-Bisphosphate Aldolase (FBA). This was supported by a NCBI BLAST which gave as top results other FBAs. Additionally, a PDB BLAST [[23]](http://198.128.27.8/genome/converter.php#ref-23) gave an article with active site residues for class II FBA. These active site residues matched with the ones in microcoleus vaginatus, except for one at position 256, where S256 became T256. The article mentioned that S256 was important for its side chain hydroxyl group. T256 would also have a side chain hydroxyl group, and serine and threonine are very similar in structure. Finally, a multiple sequence alignment using FASTA sequences from UniProt showed that for cyanobacteria, the T256 active site residue is completely conserved.

For gene 2505169777, it was also predicted to be a FBA. A NCBI BLAST also gave as top results other FBAs. A PDB BLAST [[14]](http://198.128.27.8/genome/converter.php#ref-14) gave an article about archael FBAs. The active sites from this article matched with the active sites in microcoleus vaginatus, except for site 22, where A22 was P22 for microcoleus vaginatus. The article says however that A22 is not conserved, so the enzyme from microcoleus vaginatus could be similar to archael FBAs.

**Triose-phosphate isomerase**

Gene OID: 2505168240 EC: 5.3.1.1

The gene (2505168240) is confirmed to code for the triose-phosphate isomerase enzyme (EC 5.3.1.1). A NCBI BLAST search against the swissprot database gives other triose-phosphate isomerases as top hits. A PDB BLAST [[6]](http://198.128.27.8/genome/converter.php#ref-6) search gave an article on the triose-phosophate isomerase enzyme from *Bacillus stearothermophilus*, which is a thermophilic bacteria, similar to *Microcoleus vaginatus* which can endure high desert temperatures. Additionally, all the active sites mentioned in this article matched with the ones in *Microcoleus vaginatus*.

**Functional Studies**

The functional studies for most of these genes already show up in the glycolysis report. The only two ECs are 2.7.1.2 and 1.1.1.27. The genes predicted to code for enzymes with function for ECs 2.7.1.2 and 1.1.1.27 should be checked.

There are some genes that are new in this pathway, and some of these genes need further functional studies. For example, gene 2505166819 needs further studies done on it. This gene was predicted to code for glucokinase (EC 2.7.1.2). However, BLAST searches do not give other glucokinases as hits, so it seems that the gene does not code for a glucokinase. This needs to be verified. There is another gene confirmed to code for glucokinase, so it would make sense if this gene doesn’t code for glucokinase, since there is another gene that takes care of coding for glucokinase.

Gene 2505168756 should also be further analyzed. This gene was predicted to code for malate dehydrogenase (EC 1.1.1.32), but it could also be L-lactate dehydrogenase (EC 1.1.1.27). Studies need to be done to see if the gene codes only for malate dehydrogenase, or if it codes for a bifunctional protein. An experiment should be done to see if *Microcoleus vaginatus* can produce L-lactate, because if it can, then it indicates the presence of L-lactate dehydrogenase, and this gene would be the gene that codes for it.

**Discussion**

It is predicted that *Microcoleus vaginatus* can undergo homolactic acid fermentation, since it has almost all the enzymes needed for the pathway. The only enzyme that is potentially missing is EC 1.1.1.27, L-lactate dehydrogenase. If there is no enzyme with this function in *Microcoleus vaginatus*, it may be the case that the organism uses its D-lactate dehdyrogenase instead, so that it converts pyruvate into D-lactate instead of L-lactate. The bottom line is that *Microcoleus vaginatus* has an enzyme to convert pyruvate to lactate, so it could possibly undergo the homolactic acid fermentation pathway.

An experiment should be done to see if *Microcoleus vaginatus* indeed uses this pathway. The experiment should be done under anaerobic conditions, such that it will force the organism to use this fermentation pathway. It should be observed if the organism can convert glucose to lactate.

If *Microcoleus vaginatus* does undergo fermentation and if it doesn’t have EC 1.1.1.27, this would be very interesting, because it would mean that the organism produces D-lactate instead of the usual L-lactate. Most organisms undergoing fermentation produce L-lactate, not D-lactate, so this would be something for additional future investigation.

**References**

[1] Angell S, Schwarz E, Bibb MJ. "The glucose kinase gene of Streptomyces coelicolor A3(2):

its nucleotide sequence, transcriptional analysis and role in glucose repression." Mol

Microbiol. 1992 Oct; 6(19):2833-44.

[2] Arai K, Ishimitsu T, Fushinobu S, Uchikoba H, Matsuzawa H, Taguchi H. "Active and inactive

state structures of unliganded Lactobacillus casei allosteric L-lactate dehydrogenase."

Proteins. Feb 15, 2010; 78(3):681-94.

[3] Arsenieva D, Appavu BL, Mazock GH, Jeffery CJ. "Crystal structure of phosphoglucose

isomerase from Trypanosoma brucei complexed with glucose-6-phosphate at 1.6 A

resolution." Proteins. 2009 Jan; 74(1):72-80.

[4] Auerbach G, Huber R, Grättinger M, Zaiss K, Schurig H, Jaenicke R, Jacob U. "Closed

structure of phosphoglycerate kinase from Thermotoga maritima reveals the catalytic

mechanism and determinants of thermal stability." Structure. Nov 15, 1997;

5(11):1475-83.

[5] Delarue M, Duclert-Savatier N, Miclet E, Haouz A, Giganti D, Ouazzani J, Lopez P, Nilges M,

Stoven V. "Three dimensional structure and implications for the catalytic mechanism of

6-phosphogluconolactonase from Trypanosoma brucei." J Mol Biol. Feb 23, 2007;

366(3):868-81.

[6] Delboni LF, Mande SC, Rentier-Delrue F, Mainfroid V, Turley S, Vellieux FM, Martial JA, Hol

WG. "Crystal structure of recombinant triosephosphate isomerase from Bacillus

stearothermophilus. An analysis of potential thermostability factors in six isomerases with

known three-dimensional structures points to the importance of hydrophobic

interactions." Protein Sci. 1995 Dec; 4(12):2594-604.

[7] Jedrzejas MJ, Chander M, Setlow P, Krishnasamy G. "Mechanism of catalysis of the

cofactor-independent phosphoglycerate mutase from Bacillus stearothermophilus.

Crystal structure of the complex with 2-phosphoglycerate." J Biol Chem. Jul 28, 2000;

275(30):23146-53.

[8] Kaneko T, Nakamura Y, Sato S, Asamizu E, Kato T, Sasamoto S, Watanabe A, Idesawa K,

Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kiyokawa C, Kohara M, Matsumoto M,

Matsuno A, Mochizuki Y, Nakayama S, Nakazaki N, Shimpo S, Sugimoto M, Takeuchi C,

Yamada M, Tabata S. "Complete genome structure of the nitrogen-fixing symbiotic

bacterium Mesorhizobium loti." DNA Res. Dec 31, 2000; 7(6):331-8.

[9] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[10] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[11] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo

K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M,

Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[12] Kaneko T, Tanaka A, Sato S, Kotani H, Sazuka T, Miyajima N, Sugiura M, Tabata S.

"Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp.

strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92%

of the genome." DNA Res. Aug 31, 1995; 2(4):153-66, 191-8.

[13] Kühnel K, Luisi BF. "Crystal structure of the Escherichia coli RNA degradosome component

enolase." J Mol Biol. Oct 26, 2001; 313(3):583-92.

[14] Lorentzen E, Pohl E, Zwart P, Stark A, Russell RB, Knura T, Hensel R, Siebers B. "Crystal

structure of an archaeal class I aldolase and the evolution of (betaalpha)8 barrel

proteins." J Biol Chem. Nov 21, 2003; 278(47):47253-60.

[15] Louis A, Ollivier E, Aude JC, Risler JL. "Massive sequence comparisons as a help in

annotating genomic sequences." Genome Res. 2001 Jul; 11(7):1296-303.

[16] Lucas J. Stal, Roy Moezelaar"Fermentation in cyanobacteria" FEMS Microbiology Reviews.

0; 21(2):179-211.

[17] Lunin VV, Li Y, Schrag JD, Iannuzzi P, Cygler M, Matte A. "Crystal structures of Escherichia

coli ATP-dependent glucokinase and its complex with glucose." J Bacteriol. 2004 Oct;

186(20):6915-27.

[18] Martin W, Brinkmann H, Savonna C, Cerff R. "Evidence for a chimeric nature of nuclear

genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase

genes." Proc Natl Acad Sci U S A. Sep 15, 1993; 90(18):8692-6.

[19] Mattevi A, Valentini G, Rizzi M, Speranza ML, Bolognesi M, Coda A. "Crystal structure of

Escherichia coli pyruvate kinase type I: molecular basis of the allosteric transition."

Structure. Jul 15, 1995; 3(7):729-41.

[20] Rigden DJ, Littlejohn JE, Henderson K, Jedrzejas MJ. "Structures of phosphate and

trivanadate complexes of Bacillus stearothermophilus phosphatase PhoE: structural and

functional analysis in the cofactor-dependent phosphoglycerate mutase superfamily." J

Mol Biol. Jan 17, 2003; 325(3):411-20.

[21] Schirmer T, Evans PR. "Structural basis of the allosteric behaviour of phosphofructokinase."

Nature. Jan 11, 1990; 343(6254):140-5.

[22] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[23] Shao D, Zhong X, Zhou YF, Han Z, Lin Y, Wang Z, Bu L, Zhang L, Su XD, Wang H.

"Structural and functional comparison of MIF ortholog from Plasmodium yoelii with MIF

from its rodent host." Mol Immunol. 2010 Jan; 47(4):726-37.

[24] Teplyakov A, Lim K, Zhu PP, Kapadia G, Chen CC, Schwartz J, Howard A, Reddy PT,

Peterkofsky A, Herzberg O. "Structure of phosphorylated enzyme I, the

phosphoenolpyruvate:sugar phosphotransferase system sugar translocation signal

protein." Proc Natl Acad Sci U S A. Oct 31, 2006; 103(44):16218-23.

[25] Uhrínová S, Uhrín D, Nairn J, Price NC, Fothergill-Gilmore LA, Barlow PN. "Solution structure

and dynamics of an open beta-sheet, glycolytic enzyme, monomeric 23.7 kDa

phosphoglycerate mutase from Schizosaccharomyces pombe." J Mol Biol. Feb 16, 2001;

306(2):275-90.

[26] Watkins HA, Baker EN. "Structural and functional analysis of Rv3214 from Mycobacterium

tuberculosis, a protein with conflicting functional annotations, leads to its characterization

as a phosphatase." J Bacteriol. 2006 May; 188(10):3589-99.

[27] Watkins HA, Baker EN. "Structural and functional analysis of Rv3214 from Mycobacterium

tuberculosis, a protein with conflicting functional annotations, leads to its characterization

as a phosphatase." J Bacteriol. 2006 May; 188(10):3589-99.

[28] Yun M, Park CG, Kim JY, Park HW. "Structural analysis of glyceraldehyde 3-phosphate

dehydrogenase from Escherichia coli: direct evidence of substrate binding and

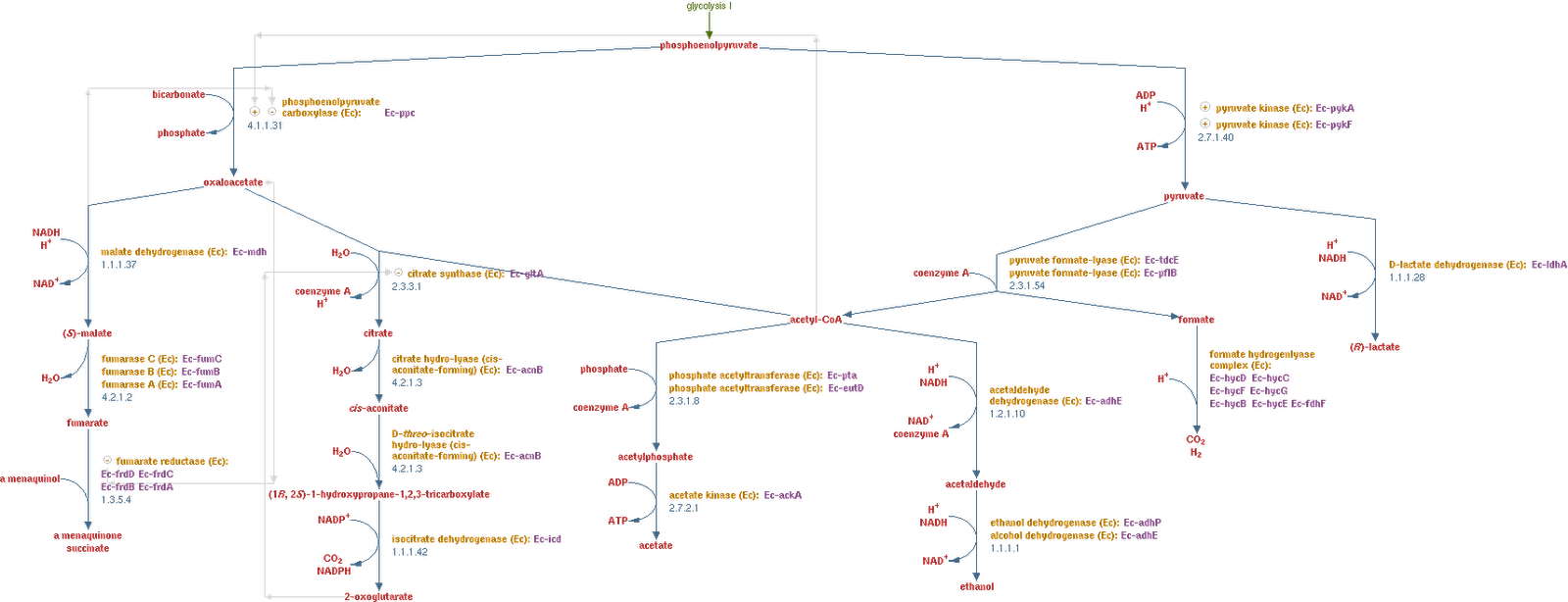
cofactor-induced conformational changes." Biochemistry. Sep 5, 2000; 39(35):10702-10.

## Mixed Acid Fermentation

Cyanobacteria are documented in the literature to undergo various types of fermentation {{DB: FERMENTATION}}. One such variation of fermentation is mixed acid fermentation.

Fermentation produces energy in an anaerobic environment via substrate level phosphorylation. The electron donor is an organic compound and the electron acceptor is another organic compound of lower energy content. In mixed acid fermentation, phosphoenolpyruvate is broken down via different branches of the pathway to a menaquinone succinate, 2-oxoglutarate, acetate, ethanol, carbon dixoide, L-lactate and molecular hydrogen. The branch that ends with acetate regenerates ATP and the branch ending with 2-oxoglutarate regenerates NADPH, while the other branches use reducing power (NADH).

**Automated Annotation**



S Fig. 1: MetaCyc map of mixed acid fermentation. Larger version at <http://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=FERMENTATION-PWY&detail-level=2&detail-level=1&detail-level=2&detail-level=1>

**Phosphoenolpyruvate carboxylase**

Gene OID: 2505170379 EC: 4.1.1.31

The gene (2505170379) was confirmed to code for EC 4.1.1.31. A NCBI BLAST (NCBI: Q3MB82) search against the swissprot database gives other Phosphoenolpyruvate carboxylase as top hits. A PDB BLAST search [[17]](http://198.128.26.117/genome/converter.php#ref-17) gave an article on the Phosphoenolpyruvate carboxylase structure, and all the active sites but one mentioned in the article matched with the ones in *Microcoleus vaginatus*.

**Malate Dehydrogenase**

Gene OID: 2505168756 EC: 1.1.1.37

Gene is confirmed to code for enzyme 1.1.1.37. A NCBI BLAST search [[9]](http://198.128.26.117/genome/converter.php#ref-9) gave other malate dehydrogenases as top hits. A PDB BLAST search [[6]](http://198.128.26.117/genome/converter.php#ref-6) gave as a top hit an article on the structure of malate dehydrogenases. From this article, the active site residues were compared, and all but one of them matched. The mismatch at site 10 involved the replacing of an Asn with a Lys. According to the article, the side chain nitrogen group Asn 10 interacts through hydrogen bonding with the two phosphate groups in one of the substrates, NAD. In microcoleus vaginatus, Lys 10 could also feasibly interact through hydrogen bonding, because Lys also has a side chain nitrogen group.

As another check, a multiple sequence alignment was done on T-Coffee, using FASTA sequences of malate dehydrogenases in other cyanobacteria . The results showed that at site 10, most cyanobacteria either had Asn or Arg as the amino acid. Some did have Lys at site 10. According to T-Coffee, the match between Asn, Arg and Lys shows up as good, and the hypotheses from earlier is confirmed because these three amino acids all have side chain nitrogen groups. It seems plausible that either Asn 10, Arg 10 or Lys 10 would be acceptable active sites for this enzyme.

**Fumarate hydratase/Aspartate Ammonia Lyase**

Gene OID: 2505169957 EC: 4.2.1.2/4.3.1.1

This gene was previously annotated as aspartate ammonia lyase, but it could very possibly code for fumarate hydratase as well. A NCBI BLAST search gave as top hits aspartate ammonia lyases [[25]](http://198.128.26.117/genome/converter.php#ref-25) with e-values as low as 8E-146 and with percent identity as high as 49%. The NCBI BLAST search also gave fumarate hydratases as top hits [[21]](http://198.128.26.117/genome/converter.php#ref-21) with e-values as low as 2E-140 and as high as 47% identity. From the NCBI BLAST, it seemed that the gene could code for either protein, since the hits were good to both aspartate ammonia lyase and fumarate hydratase.

A PDB BLAST search also resulted in the same findings. The top articles were on aspartate ammonia lyases [[7]](http://198.128.26.117/genome/converter.php#ref-7), with the top hit having an e-value of 1.66682E-116 and 46% identity. There were also articles about fumarate hydratases [[26]](http://198.128.26.117/genome/converter.php#ref-26). One of the top hit had an e-value of 3.73056E-100 and 42% identity. Active sites were mentioned in the article. For the active sites, 6 out of 7 matched, with the only mismatch being S98A. This mismatch could be important because S98 was mentioned as being important for the hydrogen bond it can form with its side chain hydroxyl group. A98 has a methyl group and thus cannot hydrogen bond. However, it was mentioned that the most important active site was H188, which matched, so it is still predicted that the gene could code for fumarate hydratase.

This gene was previously annotated for aspartate ammonia lyase function, but it probably has fumarate hydratase function as well. The protein coded for by this gene is most likely bifunctional.

**2-methylcitrate synthase/citrate synthase II**

Gene OID: 2505168309 EC: 2.3.3.1

The gene is confirmed to code for enzyme 2.3.3.1. NCBI BLAST [[13]](http://198.128.26.117/genome/converter.php#ref-13) reveals other citrate synthetases as excellent matches. Interestingly, all the top matches from PDB [[23]](http://198.128.26.117/genome/converter.php#ref-23) reveal that *Microcoleus*’ copy of this gene is very closely related to thermophilic bacteria and archaea such as *Thermus*, *Sulfolobus*, and *Pyrococcus*. All the active site residues were matches or near-matches (structurally similar amino acid replacements). This suggests that this gene is probably adapted for the extreme ground temperatures that *Microcoleus vaginatus* endures.

**Aconitase**

Gene OID: 2505168762 EC: 4.2.1.3

Gene is confirmed to code for enzyme 4.2.1.3. A NCBI BLAST search [[12]](http://198.128.26.117/genome/converter.php#ref-12) against the Swissprot database gives other aconitases as top hits. A PDB BLAST search [[27]](http://198.128.26.117/genome/converter.php#ref-27) gave an article on E. coli aconitase structure, and all the active sites mentioned in this article matched with the ones in *Microcoleus vaginatus*.

**Isocitrate Dehydrogenase (NADP-dependent)**

Gene OID: 2505168569 EC: 1.1.1.42

Gene is confirmed to code for enzyme 1.1.1.42. A NCBI BLAST search [[20]](http://198.128.26.117/genome/converter.php#ref-20) against the Swissprot database gives other isocitrate dehydrogenases as top hits. A PDB BLAST search [[19]](http://198.128.26.117/genome/converter.php#ref-19) gave an article on isocitrate dehydrogenase structure, and all but one of the active sites mentioned in this article matched with the ones in *Microcoleus vaginatus*. The article says the active site at 345 should be a tyrosine, but for *Microcoleus vaginatus* it is a histidine. A multiple sequence alignment was performed on T-Coffee, which showed the histidine residue to be completely conserved at site 345 for other cyanobacteria. The FASTA sequences used for comparison were obtained from UniProt.

**Phosphate acetyltransferase**

EC: 2.3.1.8

No gene was found in *Microcoleus vaginatus* that codes for this enzyme. IMG/ER predicted no genes to have function for this enzymes. A search was done on UniProt for cyanobacterial enzymes with EC 2.3.1.8 function. Their FASTA sequences were BLASTed against the *Microcoleus vaginatus* genome in IMG/ER, but no strong hits were obtained. The hits from the BLAST search had e-values and didn’t correspond to phosphate acetyltransferases.

**Acetate Kinase**

Gene OID: 2505165972 EC: 2.7.2.1

The gene is predicted to code for acetate kinase. A NCBI BLAST search gave as top hits other acetate kinases [[10]](http://198.128.26.117/genome/converter.php#ref-10) as top hits, with e-values as low as 0 and percent identities as high as 68%. A PDB BLAST search gave articles on acetate kinases as top hits, with the best hit having an e-value of 4.72107E-85 and 42% identity [[4]](http://198.128.26.117/genome/converter.php#ref-4). 17 active sites were mentioned in the article, and all 17 were matched. A reverse BLAST search gives this gene as the first hit. This gene is therefore strongly predicted to code for acetate kinase.

**Pyruvate Kinase**

Gene OID: 2505167661 EC: 2.7.1.40

Gene is predicted to code for pyruvate kinase. A NCBI BLAST returned a near-perfect match with another pyruvate kinase, and gave as top results other pyruvate kinases. The top hit had an e-value of 0 and a query coverage of 98% [[11]](http://198.128.26.117/genome/converter.php#ref-11). A reverse BLAST search also gave the gene as a top hit, with an e-value of 0. A PDB BLAST search gave an article on pyruvate kinase structure in E. coli. This particular hit had an e-value of 9.39729E-98 and a query coverage of 42%. All the active sites mentioned in this article matched with the active sites from *Microcoleus vaginatus*[[18]](http://198.128.26.117/genome/converter.php#ref-18).

**D-Lactate Dehydrogenase**

Gene OID: 2505170579 EC: 1.1.1.28

This gene is predicted to code for D-lactate dehydrogenase. A NCBI BLAST search gave D-lactate dehydrogenases as top hits [[3]](http://198.128.26.117/genome/converter.php#ref-3) with e-values from 1E-126 to 1E-44 and around 40% to 50% identity. A PDB BLAST search gave an article on D-lactate dehdyrogenase as a top hit[[22]](http://198.128.26.117/genome/converter.php#ref-22) with an e-value of 9.60815E-49 and 35% identity. The article mentioned active sites, and 8 out 11 of these matched. The mismatches were V233T, H155K and N212T. The article mentioned V233 being important to to its hydrophobicity, but T233 would not be hydrophobic. The article also mentioned H155 being important for its imidazole ring to interact with bridging water molecules. K155 doesn’t have an imidazole ring, but it also has a side chain nitrogen that could similar interact with bridging water molecules. N212 was mentioned to be important for participating in 2 hydrogen bonds, which T212 could do as well, since part of its side chain (the NH2 group) can participate in hydrogen bonding. As further confirmation, a multiple sequence alignment was done on T-Coffee with cyanobacterial FASTA sequences from UniProt. This showed that at site 155, no other cyanobacteria had H155, and all but one of them had the K155 residue. At site 212, the N212 residue was not found among any other cyanobacteria, while all but one cyanobacteria had the T212 residue. At site V233, no other cyanobacteria had residue V233, while all but one had residue T233. All these considerations taken together leads to the conclusion that the gene probably codes for D-lactate dehydrogenase, along with the fact that a reverse BLAST gave the same gene as the top hit.

**Alcohol dehydrogenase**

Gene OID: 2505166244 EC: 1.1.1.1

Gene is predicted to code for alcohol dehydrogenase. A NCBI BLAST search gave other alcohol dehydrogenases as top hits[[1]](http://198.128.26.117/genome/converter.php#ref-1) with e-values from 0 to 2E-100 and around 60% to 70% identity. A PDB BLAST search gave an article on alcohol dehydrogenase[[24]](http://198.128.26.117/genome/converter.php#ref-24) with an e-value of 4.76389E-137 and 63% identity. All the active sites mentioned in the article matched, and a reverse BLAST search gave gene 2505166244 as the top hit with an e-value of 0. Therefore, the gene is strongly predicted to code for alcohol dehydrogenase.

**Acetaldehyde Dehydrogenase**

EC: 1.2.1.10

No gene is predicted in *Microcoleus vaginatus* to code for this enzyme. IMG/ER did not automatically predict a gene for this protein. A search was done on UniProt for cyanobacterial acetaldehyde dehydrogenase FASTA sequences, and these sequences were BLASTed against the *Microcoleus vaginatus* genome in IMG/ER. There were no strong hits. The hits all had high e-values, and there were no acetaldehyde dehydrogenase genes.

**Pyruvate Formate-Lyase**

EC: 2.3.1.54

No gene is predicted in *Microcoleus vaginatus* to code for this enzyme. IMG/ER did not automatically predict a gene for this protein. A search was done on UniProt for cyanobacterial pyruvate formate-lyase FASTA sequences, and these sequences were BLASTed against the *Microcoleus vaginatus* genome in IMG/ER. There were no strong hits. The hits all had high e-values, and there were no pyruvate formate-lyase genes.

**Fumarate Reductase**

EC: 1.3.5.4

This EC number corresponded to a complex, not just a single enzyme. This complex has 4 subunits, and bears close homology to succinate dehydrognease. Two of the subunits, FrdA and FrdB, had genes found for them. The other two subunits, FrdC and FrdD, did not. As a whole, the fumarate reductase complex isn’t predicted to be in *Microcoleus vaginatus* because two of the subunits are missing.

Subunit FrdA

Gene OID: 2505168142

Gene is predicted to code for subunit FrdA of the fumarate reductase complex. A NCBI BLAST gave succinate dehydrogenase flavoproteins as top hits, and also fumarate reductase flavoprotein subunits [[5]](http://198.128.26.117/genome/converter.php#ref-5), with the top hit having an e-value of 1E-140 and 41% identity. A PDB BLAST was also done, and this gave articles on succinate dehydrogenase, and also on fumarate reductase [[16]](http://198.128.26.117/genome/converter.php#ref-16), with this hit having an e-value of

6.05691E-116 and 40% identity. Active sites were mentioned in this article, and 5 out of 6 matched. The mismatch was T244S, which is not a grave mismatch. T244 and S244 are very similar in structure. S244 only has an extra carbon chain on its side chain; both of them have a hydroxyl group and can participate in hydrogen bonding. Therefore, this gene is predicted to code for the FrdA subunit.

Subunit FrdB

Gene OID: 2505166800

Gene is predicted to code for the FrdB subunit of the fumarate reductase complex. A NCBI BLAST search gave succinate dehydrogenase iron-sulfur subunits as top hits, and then fumarate reductase iron-sulfur subunits as well {{NCBI: P20921}, with a low e-value of 2E-45 and a high percent identity of 37%. A PDB BLAST search gave articles on succinate dehydrogenase as top hits, but also gave articles on fumarate reductase [[8]](http://198.128.26.117/genome/converter.php#ref-8), with a low e-value of 5.00974E-37 and 36% identity. A reverse BLAST search also gave back the same gene. From the strength of these hits, the gene is predicted to code for the FrdB subunit.

Subunit FrdC

No gene was predicted to code for this subunit, which codes for part of the fumarate reductase membrane protein. A IMG/ER BLAST against the *Microcoleus vaginatus* genome with FrdC FASTA sequences from UniProt gave no good hits. Additionally, no other cyanobacteria were found to have the FrdC protein on UniProt, so *Microcoleus vaginatus* is not unique in not having a FrdC subunit.

Subunit FrdD

No gene was predicted to code for this subunit, which codes for part of the fumarate reductase membrane protein. A IMG/ER BLAST against the *Microcoleus vaginatus* genome with FrdD FASTA sequences from UniProt gave no good hits. Additionally, no other cyanobacteria were found to have the FrdD protein on UniProt, so *Microcoleus vaginatus* is not unique in not having a FrdD subunit.

**Formate hydrogenlyase complex**

This complex does not have an EC number assigned to it. The complex consists of 7 subunits. The only subunit found to have a gene coding for it in *Microcoleus vaginatus* was the fdhF subunit. Therefore, it is predicted that *Microcoleus vaginatus* does not have a formate hydrogenlyase complex, since it is missing six of the subunits necessary for the complex.

Subunit hycB

No gene was predicted to code for this subunit, which codes for part of the fumarate reductase membrane protein. A IMG/ER BLAST against the *Microcoleus vaginatus* genome with hycB FASTA sequences from UniProt gave no good hits. Additionally, no other cyanobacteria were found to have the hycB protein on UniProt, so *Microcoleus vaginatus* is not unique in not having a hycB subunit.

Subunit hycC

No gene was predicted to code for this subunit, which codes for part of the fumarate reductase membrane protein. A IMG/ER BLAST against the *Microcoleus vaginatus* genome with hycC FASTA sequences from UniProt gave no good hits. The hits with decent e-values were for ubiquinone oxidoreductase subunit.

Subunit hycD

No gene was predicted to code for this subunit, which codes for part of the fumarate reductase membrane protein. A IMG/ER BLAST against the *Microcoleus vaginatus* genome with hycD FASTA sequences from UniProt gave no good hits. The hits with decent e-values were for ubiquinone oxidoreductase subunit.

Subunit hycE

No gene was predicted to code for this subunit, which codes for part of the fumarate reductase membrane protein. A IMG/ER BLAST against the *Microcoleus vaginatus* genome with hycE FASTA sequences from UniProt gave no good hits. The hits with decent e-values were for ubiquinone oxidoreductase subunit.

Subunit hycF

No gene was predicted to code for this subunit, which codes for part of the fumarate reductase membrane protein. A IMG/ER BLAST against the *Microcoleus vaginatus* genome with hycF FASTA sequences from UniProt gave no good hits. The hits with decent e-values were for ubiquinone oxidoreductase subunit.

Subunit hycG

No gene was predicted to code for this subunit, which codes for part of the fumarate reductase membrane protein. A IMG/ER BLAST against the *Microcoleus vaginatus* genome with hycG FASTA sequences from UniProt gave no good hits. The hits with decent e-values were for ubiquinone oxidoreductase subunit.

Subunit fdhF

Gene OID: 2505171113

The gene was predicted to code for subunit fdhF. A NCBI BLAST search gave some hits to putative formate dehydrogenases [[14]](http://198.128.26.117/genome/converter.php#ref-14) with e-values around 1E-91 and around 29% identity. A PDB BLAST search gave a hit to an article about formate dehydrogenase [[2]](http://198.128.26.117/genome/converter.php#ref-2) with an e-value of 3.50518E-96 and 30% identity. Out of 12 active sites mentioned, 10 of them matched. The mismatches were G402A and D429E. G402 and A402 are very similar. The only difference is that A402 has a methyl group instead of a hydrogen, but both are nonpolar. D429 and E429 are also very similar, with E429 just having an extra carbon branch on its side chain. Besides that, everything else is the same. Due to these facts, the gene is predicted to code for subunit fdhF.

**Functional Studies**

Many of the proteins and the genes that code for the proteins in this pathway have already been annotated in the report on the TCA cycle. Functional studies for those proteins were mentioned in the TCA cycle report. There were also some genes that were not found that were already mentioned to be missing in the heterolactic acid fermentation pathway as well. For this mixed acid fermentation pathway, the new proteins were fumarate reductase, formate dehydrogenase and pyruvate formate-lyase.

Fumarate reductase is a protein complex with 4 subunits, and genes were predicted to code for only 2 out of the 4 subunits of this complex. Therefore, fumarate reductase was not predicted to be found in *Microcoleus vaginatus*. The reason the 2 subunits are there is most likely because they code for subunits to succinate dehydrogenase, which bears strong homology to fumarate reductase. A study could be done to confirm if the organism is missing this complex, by observing if *Microcoleus vaginatus* can convert fumarate to a menaquinone succinate.

Formate dehydrogenase is also a protein complex, but it has 7 subunits. Only one subunit was predicted to be coded for by a gene in *Microcoleus vaginatus*. Therefore, the protein complex was not predicted as a whole to exist in the organism. To double-check, and experiment could be done to see if *Microcoleus vaginatus* can convert formate to carbon dixoide and hydrogen.

Finally, no gene was found to code for the pyruvate formate-lyase enzyme. This prediction could also be tested with an experiment. In the experiment, it should be seen if *Microcoleus vaginatus* can convert pyruvate to formate and acetyl CoA, by observing if increasing pyruvate levels under anaerboic conditions increases formate and acetyl CoA levels.

**Discussion**

*Microcoleus vaginatus* is not predicted to be able to undergo the mixed acid fermentation pathway, because it is missing too many enzymes necessary for the pathway. It could however theoretically undergo a modified version of this pathway. On the left branch of the pathway shown in S Figure 1, *Microcoleus vaginatus* could convert S-malate to fumarate and stop there, not converting fumarate to a menaquinone succinate because it is missing the necessary protein complex. Along the right side, since it is missing pyruvate formate-lyase, *Microcoleus vaginatus* cannot convert pyruvate to acetyl CoA and formate. Since it cannot convert pyruvate to formate and there is no alternate path to produce formate, the lack of the formate dehydrogenase complex becomes a moot point. *Microcoleus vaginatus* would simply convert all its pyruvate to R-lactate. As far as acetyl CoA is concerned, *Microcoleus vaginatus* can still produce acetyl CoA via oxaloacetate.

In conclusion, *Microcoleus vaginatus* is known to undergo other fermentation pathways and has enough proteins that it could undergo parts of this mixed acid fermentation pathway. However, further research needs to be done in the lab though to see which fermentation pathway *Microcoleus vaginatus* actually uses in real life.

**References**

[1] Barber RD, Rott MA, Donohue TJ. "Characterization of a glutathione-dependent formaldehyde

dehydrogenase from Rhodobacter sphaeroides." J Bacteriol. 1996 Mar; 178(5):1386-93.

[2] Boyington JC, Gladyshev VN, Khangulov SV, Stadtman TC, Sun PD. "Crystal structure of

formate dehydrogenase H: catalysis involving Mo, molybdopterin, selenocysteine, and an

Fe4S4 cluster." Science. Feb 28, 1997; 275(5304):1305-8.

[3] Bunch PK, Mat-Jan F, Lee N, Clark DP. "The ldhA gene encoding the fermentative lactate

dehydrogenase of Escherichia coli." Microbiology. 1997 Jan; 143 ( Pt 1)():187-95.

[4] Buss KA, Cooper DR, Ingram-Smith C, Ferry JG, Sanders DA, Hasson MS. "Urkinase:

structure of acetate kinase, a member of the ASKHA superfamily of

phosphotransferases." J Bacteriol. 2001 Jan; 183(2):680-6.

[5] Cole ST. "Nucleotide sequence and comparative analysis of the frd operon encoding the

fumarate reductase of Proteus vulgaris. Extensive sequence divergence of the

membrane anchors and absence of an frd-linked ampC cephalosporinase gene." Eur J

Biochem. Sep 15, 1987; 167(3):481-8.

[6] Dalhus B, Saarinen M, Sauer UH, Eklund P, Johansson K, Karlsson A, Ramaswamy S, Bjørk

A, Synstad B, Naterstad K, Sirevåg R, Eklund H. "Structural basis for thermophilic protein

stability: structures of thermophilic and mesophilic malate dehydrogenases." J Mol Biol.

May 3, 2002; 318(3):707-21.

[7] Fujii T, Sakai H, Kawata Y, Hata Y. "Crystal structure of thermostable aspartase from Bacillus

sp. YM55-1: structure-based exploration of functional sites in the aspartase family." J Mol

Biol. May 2, 2003; 328(3):635-54.

[8] Iverson TM, Luna-Chavez C, Croal LR, Cecchini G, Rees DC. "Crystallographic studies of

the Escherichia coli quinol-fumarate reductase with inhibitors bound to the quinol-binding

site." J Biol Chem. May 3, 2002; 277(18):16124-30.

[9] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[10] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete

genomic sequence of the filamentous nitrogen-fixing cyanobacterium Anabaena sp.

strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[11] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo

K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M,

Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[12] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo

K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M,

Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[13] Kaneko T, Tanaka A, Sato S, Kotani H, Sazuka T, Miyajima N, Sugiura M, Tabata S.

"Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp.

strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92%

of the genome." DNA Res. Aug 31, 1995; 2(4):153-66, 191-8.

[14] Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, Cui L, Oguchi A, Aoki K,

Nagai Y, Lian J, Ito T, Kanamori M, Matsumaru H, Maruyama A, Murakami H, Hosoyama

A, Mizutani-Ui Y, Takahashi NK, Sawano T, Inoue R, Kaito C, Sekimizu K, Hirakawa H,

Kuhara S, Goto S, Yabuzaki J, Kanehisa M, Yamashita A, Oshima K, Furuya K, Yoshino

C, Shiba T, Hattori M, Ogasawara N, Hayashi H, Hiramatsu K. "Whole genome

sequencing of meticillin-resistant Staphylococcus aureus." Lancet. Apr 21, 2001;

357(9264):1225-40.

[15] Lucas J. Stal, Roy Moezelaar "Fermentation in cyanobacteria" FEMS Microbiology Reviews.

0; 21(2):179-211.

[16] Maklashina E, Iverson TM, Sher Y, Kotlyar V, Andréll J, Mirza O, Hudson JM, Armstrong FA,

Rothery RA, Weiner JH, Cecchini G. "Fumarate reductase and succinate oxidase activity

of Escherichia coli complex II homologs are perturbed differently by mutation of the flavin

binding domain." J Biol Chem. Apr 21, 2006; 281(16):11357-65.

[17] Matsumura H, Xie Y, Shirakata S, Inoue T, Yoshinaga T, Ueno Y, Izui K, Kai Y. "Crystal

structures of C4 form maize and quaternary complex of E. coli phosphoenolpyruvate

carboxylases." Structure. 2002 Dec; 10(12):1721-30.

[18] Mattevi A, Valentini G, Rizzi M, Speranza ML, Bolognesi M, Coda A. "Crystal structure of

Escherichia coli pyruvate kinase type I: molecular basis of the allosteric transition."

Structure. Jul 15, 1995; 3(7):729-41.

[19] Mesecar AD, Stoddard BL, Koshland DE Jr. "Orbital steering in the catalytic power of

enzymes: small structural changes with large catalytic consequences." Science. Jul 11,

1997; 277(5323):202-6.

[20] Muro-Pastor MI, Florencio FJ. "NADP(+)-isocitrate dehydrogenase from the cyanobacterium

Anabaena sp. strain PCC 7120: purification and characterization of the enzyme and

cloning, sequencing, and disruption of the icd gene." J Bacteriol. 1994 May;

176(9):2718-26.

[21] Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS,

Thorsson V, Sbrogna J, Swartzell S, Weir D, Hall J, Dahl TA, Welti R, Goo YA,

Leithauser B, Keller K, Cruz R, Danson MJ, Hough DW, Maddocks DG, Jablonski PE,

Krebs MP, Angevine CM, Dale H, Isenbarger TA, Peck RF, Pohlschroder M, Spudich JL,

Jung KW, Alam M, Freitas T, Hou S, Daniels CJ, Dennis PP, Omer AD, Ebhardt H, Lowe

TM, Liang P, Riley M, Hood L, DasSarma S. "Genome sequence of Halobacterium

species NRC-1." Proc Natl Acad Sci U S A. Oct 24, 2000; 97(22):12176-81.

[22] Razeto A, Kochhar S, Hottinger H, Dauter M, Wilson KS, Lamzin VS. "Domain closure,

substrate specificity and catalysis of D-lactate dehydrogenase from Lactobacillus

bulgaricus." J Mol Biol. Apr 19, 2002; 318(1):109-19.

[23] Russell RJ, Ferguson JM, Hough DW, Danson MJ, Taylor GL. "The crystal structure of

citrate synthase from the hyperthermophilic archaeon pyrococcus furiosus at 1.9 A

resolution,." Biochemistry. Aug 19, 1997; 36(33):9983-94.

[24] Sanghani PC, Robinson H, Bosron WF, Hurley TD. "Human glutathione-dependent

formaldehyde dehydrogenase. Structures of apo, binary, and inhibitory ternary

complexes." Biochemistry. Sep 3, 2002; 41(35):10778-86.

[25] Sun DX, Setlow P. "Cloning, nucleotide sequence, and expression of the Bacillus subtilis

ans operon, which codes for L-asparaginase and L-aspartase." J Bacteriol. 1991 Jun;

173(12):3831-45.

[26] Weaver T, Lees M, Banaszak L. "Mutations of fumarase that distinguish between the active

site and a nearby dicarboxylic acid binding site." Protein Sci. 1997 Apr; 6(4):834-42.

[27] Williams CH, Stillman TJ, Barynin VV, Sedelnikova SE, Tang Y, Green J, Guest JR, Artymiuk

PJ. "E. coli aconitase B structure reveals a HEAT-like domain with implications for

protein-protein recognition." Nat Struct Biol. 2002 Jun; 9(6):447-52.

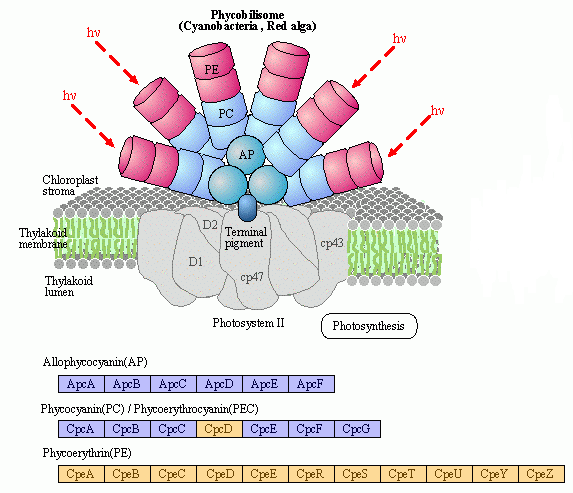
# Photosynthesis

## Antennae Protein Photosynthesis Complex

The antennae protein complex of cyanobacteria is also called a phycobilisome. Phycobilisomes are light harvesting antennae of photosystem II, anchored to the thylakoid membrane. They are made of stacks of chromophorylated proteins, the phycobiliproteins, and their associated linker polypeptides. Each phycobilisome consists of a core made of allophycocyanin, from which several outwardly oriented rods made of stacked disks of phycocyanin and (if present) phycoerythrin(s) or phycoerythrocyanin.

Each phycobilisome has a specific absorption and fluorescence emission maximum in the visible range of light. The range of light the phycobilisome can absorb differs from that of chlorophyll. In this way, the cells take advantage of the available wavelengths of light which are inaccessible to chlorophyll, and utilize this energy for photosynthesis.

**Automated Annotation**



S Fig. 1: KEGG illustration of proteins that make up the phycobilisome complex. Blue boxes indicate proteins predicted to have genes for them in *Microcoleus vaginatus.* Orange boxes indicate missing proteins.

**Allophycocyanin (AP)**

**Allophycocyanin Subunit A (ApcA)**

Gene OID: 2505166708

This gene is predicted to code for the A subunit of allophycocyanin. A NCBI BLAST search gave other allophycocyanin A subunits as top hits [[3]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-3) with e-values from 1E-102 to 1E-80 and 67% to 87% identity. A PDB BLAST search gave as the top hit an article on allophycocyanin crystal structure [[13]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-13) with an e-value of 6.1455E-80 and 86% identity. From the strength of these BLAST hits, it is predicted that this gene codes for the A subunit.

**Allophycocyanin Subunit B (ApcB)**

Gene OID: 2505166709

This gene is predicted to code for the B subunit of allophycocyanin. A NCBI BLAST search gave other allophycocyanin B subunits as top hits [[3]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-3), with e-values from 3E-107 to 5E-39 and 49% to 93% identity. A PDB BLAST search gave as the top hit an article on the crystal structure of allophycocyanin {{PDB: 1ALL}} with an e-value of 7.90031E-77

and 86% identity. From the strength of these BLAST hits, it is predicted that this gene codes for the B subunit.

**Allophycocyanin Core Linker Protein (ApcC)**

Gene OID: 2505166710

This gene is predicted to code for the allophycocyanin core linker protein. A NCBI BLAST search gave other allophycocyanin core linker proteins as top hits [[6]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-6) with e-values from 2E-42 to 3E-33 and 88% to 93% identity. A PDB BLAST search gave as the top hit an article on allophycocoyanin crystal structure [[15]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-15) with an e-value of 6.03815E-29 and 86% identity. From the strength of these BLAST hits, it is predicted that this gene codes for the core linker protein.

**Allophycocyanin-B (ApcD)**

Gene OID: 2505167161

This gene is predicted to code for allophycocyanin-B. A NCBI BLAST search gave allophycocyanin-B as top hits [[18]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-18), with e-values from 5E-93 to 7E-67 and 57% to 78% identity. A PDB BLAST search gave as the top hit an article on allophycocyanin crystal structure [[13]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-13) with an e-value of 9.32483E-52 and 57% identity. This gene is therefore predicted to code for allophycocyanin-B.

**Allophycocyanin Core-Membrane Linker Protein (ApcE)**

Gene OID: 2505169610

This gene is predicted to code for the allophycocyanin core-membrane linker protein. A NCBI BLAST search gave allophycocyanin core-membrane linker proteins as top hits [[8]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-8) with e-values around 0 and around 65% to 72% identity. A PDB BLAST search gave many hits to unpublished article. The top hit that was published was an article on allophycocyanin crystal structure [[12]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-12). This hit had an e-value of 5.5225E-6 and 38% identity. The e-value was low because this was around the 10th top hit. The other top hits, which were unpublished, had much higher e-values. In conclusion, this gene is predicted to code for the allophycocyanin core-membrane linker protein.

**Allophycocyanin Core Component (ApcF)**

Gene OID: 2505167134

This gene is predicted to code for the allophycocyanin core component. A NCBI BLAST search gave allophycocyanin subunit beta-18 hits (another name for allophycocyanin core component) as top hits [[7]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-7) with e-values from 2E-85 to 1E-62 and 51% to 78% identity. A PDB BLAST search gave as a top hit an article on allophycocyanin {{PDB: 1ALL}} with an e-value of 2.07017E-36 and 49% identity. The hits from NCBI and PDB are strong, so this gene is predicted to code for allophycocyanin core component.

**Phycocyanin (PC) / Phycoerythrocyanin (PEC)**

**Phycocyanin Alpha Subunit (CpcA)**

Gene OID: 2505169787

This gene is predicted to code for phycocyanin alpha subunit. A NCBI BLAST search gave hits to phycocyanin alpha subunits [[1]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-1), with e-values from 3E-104 to 3E-82 with 75% to 86% identity. A PDB BLAST search gave as a top hit an article on phycocyanin crystal structure [[14]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-14) with an e-value of 1.83261E-74 and 80% identity. From the strength of these BLAST hits, this gene is predicted to code for phycocyanin alpha subunit.

**Phycocyanin Beta Subunit (CpcB)**

Gene OID: 2505169788

This gene is predicted to code for the phycocyanin beta subunit. A NCBI BLAST search gave hits to phycocyanin beta subunits [[19]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-19), with e-values ranging from 9E-104 to 2E-43 and 46% to 83% identity. A PDB BLAST search gave as a top hit an article on phycocyanin crystal structure [[17]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-17) with an e-value of 2.82606E-78 and 78% identity. From the strength of these BLAST hits, it is predicted that the gene codes for phycocyanin beta subunit.

**Phycocyanin-associated rod linker protein (CpcC)**

Two genes were predicted by IMG/ER to code for this protein. Both of them are predicted to code for the phycocyanin-associated rod linker protein.

Gene OID: 2505169785

This gene is predicted to code for phycocyanin-associated rod linker protein. A NCBI BLAST search gave phycocyanin-associated rod linker proteins as top hits [[9]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-9) with e-values from 1E-145 to 2E-64 and 44% to 70% identity. A PDB BLAST search gave hits to phycocyanin-associated rod linker proteins, but all of these hits were unpublished articles. However, just from the strength of the NCBI BLAST hits, it is predicted that this gene codes for phycocyanin-associated rod linker proteins.

Gene OID: 2505169786

This gene is predicted to code for phycocyanin-associated rod linker protein as well. A NCBI BLAST search gave phycocyanin-assoicated rod linker proteins as top hits [[9]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-9) with e-values that ranged from 1E-119 to 3E-63, and with 41% to 63% identity. As with the previous gene that was annotated, a PDB BLAST search gave hits to articles on phycocyanin-associated rod linker proteins, but all of these articles were unpublished articles. Therefore, solely on the strength of the NCBI BLAST hits, this gene is predicted to code for phycocyanin-associated rod linker protein.

**Phycocyanin-associated rod (CpcD)**

Gene OID: 2505169784

This gene is predicted to code for the phycocyanin-associated rod protein. A NCBI BLAST search gave as top hits other phycocyanin rod proteins [[21]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-21) with e-values ranging from 6E-30 to 3E-17 and 46% to 62% identity. A PDB BLAST search gave no good hits to articles on phycocyanin rod proteins. Despite this, the gene is predicted to code for the rod protein on the basis of the NCBI BLAST hits.

**Phycocyanin Lyase Alpha Subunit (CpcE)**

Gene OID: 2505169783

This gene is predicted to code for the phycocyanin lyase alpha subunit. A NCBI BLAST search gave as top hits other phycocyanin lyase alpha subunits [[2]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-2) with e-values from 1E-114 to 2E-90 and 57% to 67% identity. A PDB BLAST search gave no good hits to articles on phycocyanin lyase alpha subunits. Despite this, the gene is still predicted to code for the lyase alpha subunit on the the basis of the NCBI BLAST hits.

**Phycocyanin Lyase Beta Subunit (CpcF)**

Gene OID: 2505169782

This gene is predicted to code for the phycocyanin lyase beta subunit. A NCBI BLAST search gave as the top hits other phycocyanin lyase beta subunits [[20]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-20) with e-values ranging from 6E-66 to 5E-46 and 48% to 62% idntity. A PDB BLAST search gave no good hits to articles on phycocyanin lyase beta subunits. Despite this, the gene is still predicted to code for the lyase beta subunit, due to the NCBI BLAST hits.

**Phycobilisome Rod-Core Linker Protein (CpcG)**

Gene OID: 2505166633

This gene is predicted to code for the phycocyanin rod-core linker protein. A NCBI BLAST search gave as top hits other phycobilisome rod-core linker proteins [[4]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-4) with e-values from 6E-102 to 7E-57 and 43% to 61% identity. A PDB BLAST search gave hits to articles on the crystal structure of phycocyanin rod-core linker proteins, but none of these articles were published. Therefore, the gene is predicted to code for the phycocyanin rod-core linker protein solely on the basis of the NCBI BLAST hits.

**Phycoerythrin (PE)**

**Phycoerythrin Beta Chain (CpeB)**

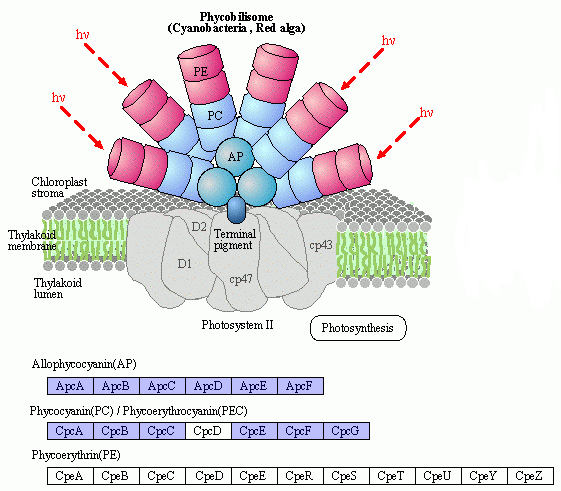
Gene OID: 2505167873

This gene is predicted to possibly code for the phycoerythrin beta chain. A NCBI BLAST search gave hits to phycoerythrin beta chain proteins [[16]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-16) with e-values from 2E-09 to 6E-07 and 26% to 29% identity. A PDB BLAST search gave hits to articles on phycoerythrin crystal structure, with the best hit having an e-value of 3.51272E-8 and 28% identity [[5]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-5). These gene could possibly code for the beta chain, but it is unsure due to the weak NCBI and PDB BLAST hits, along with the fact that a reverse BLAST search does not give this gene as the top hit.

**Genes CpeA, CpeC, CpeD, CpeE, CpeR, CpeS, CpeT, CpeU, CpeY and CpeZ**

These genes were not found in Microcoleus vaginatus. IMG/ER gave no genes as predictions for these genes, and reverse BLAST searches gave no good hits for candidate genes. The reverse BLAST searches were done using cyanobacterial FASTA sequences from the UniProt database. It seems that Microcoleus vaginatus is missing the phycoerythrin component of the phycobilisome.

**Comparison with Other Genomes**



S Fig. 2: KEGG illustration of antennae protein complex. The blue-colored boxes indicate that the gene was found in all cyanobacteria in IMG/ER, while a white box means that no genes were found in any of the cyanobacteria.

A comparison was done with other cyanobacterial genomes to see if other cyanobacteria were missing genes for the phycoeryhtrin component of the phycobilisiome. This comparison showed that all other cyanobacteria were predicted to not have the genes to code for the phycoerythrin, in addition to being predicted to not have the CpcD gene. These results were exactly the same as for *Microcoleus vaginatus*; the bug is not unique or special but rather is similar to other cyanobacteria.

**Discussion**

A literature search was done to see if phycoerythrin was found in other cyanobacteria. It was found that some cyanobacteria did express the phycoerythrin protein [[11]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-11). It does not seem like all cyanobacteria express the phycoerythrin protein, however. The phycoerythrin protein was mentioned to be a red-pigmented protein, while allophycocyanin is a blue-green pigmented protein and phycocyanin is a blue-pigmented protein [[11]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-11). It seems that these different proteins are best suited for absorbing certain wavelengths of light, such that a cyanobacteria would express the proteins corresponding to the wavelenghts of light it usually receives. It is noted in the literature that marine cyanobacteria, such as those that live in lakes, will express high levels of phycoerythrin because this protein can efficiently absorb the blue/green light that prevails in these types of waters [[10]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-10).

Microcoleus vaginatus is not a such a marine cyanobacteria. It does not live in lakes or bodies of water, so it would make sense for it to not have phycoerythrin, since this is a protein that is more efficient for absorbing light that is passing through water and is turned into hues of blue and green. Nevertheless, it would be interesting to test this hypothesis in the lab, and see if Microcoleus vaginatus is indeed not producing phycoerythrin. This could be observed by microscopy, and it could be seen visually whether or not the phycoerythrin protein structure is present. Experiments could also be done to see if Microcoleus vaginatus can effectively absorb wavelengths of light that phycoerythrin is optimal at absorbing.

**References**

[1] Anderson LK, Grossman AR. "Genes for phycocyanin subunits in Synechocystis sp. strain

PCC 6701 and assembly mutant UV16." J Bacteriol. 1990 Mar; 172(3):1289-96.

[2] Bhalerao RP, Lind LK, Gustafsson P. "Cloning of the cpcE and cpcF genes from

Synechococcus sp. PCC 6301 and their inactivation in Synechococcus sp. PCC 7942."

Plant Mol Biol. 1994 Oct; 26(1):313-26.

[3] Brejc K, Ficner R, Huber R, Steinbacher S. "Isolation, crystallization, crystal structure

analysis and refinement of allophycocyanin from the cyanobacterium Spirulina platensis

at 2.3 A resolution." J Mol Biol. Jun 2, 1995; 249(2):424-40.

[4] Bryant DA, de Lorimier R, Guglielmi G, Stevens SE Jr. "Structural and compositional

analyses of the phycobilisomes of Synechococcus sp. PCC 7002. Analyses of the

wild-type strain and a phycocyanin-less mutant constructed by interposon mutagenesis."

Arch Microbiol. 1990; 153(6):550-60.

[5] Chang WR, Jiang T, Wan ZL, Zhang JP, Yang ZX, Liang DC. "Crystal structure of

R-phycoerythrin from Polysiphonia urceolata at 2.8 A resolution." J Mol Biol. Oct 11,

1996; 262(5):721-31.

[6] DiMagno L, Haselkorn R. "Isolation and characterization of the genes encoding

allophycocyanin subunits and two linker proteins from Synechocystis 6714." Plant Mol

Biol. 1993 Mar; 21(5):835-45.

[7] Dong C, Tang A, Zhao J, Mullineaux CW, Shen G, Bryant DA. "ApcD is necessary for

efficient energy transfer from phycobilisomes to photosystem I and helps to prevent

photoinhibition in the cyanobacterium Synechococcus sp. PCC 7002." Biochim Biophys

Acta. 2009 Sep; 1787(9):1122-8.

[8] Ducret A, Sidler W, Wehrli E, Frank G, Zuber H. "Isolation, characterization and electron

microscopy analysis of a hemidiscoidal phycobilisome type from the cyanobacterium

Anabaena sp. PCC 7120." Eur J Biochem. Mar 15, 1996; 236(3):1010-24.

[9] Füglistaller P, Suter F, Zuber H. "Linker polypeptides of the phycobilisome from the

cyanobacterium Mastigocladus laminosus: amino-acid sequences and relationships."

Biol Chem Hoppe Seyler. 1985 Oct; 366(10):993-1001.

[10] Jasser I, Królicka A, Karnkowska-Ishikawa A. "A novel phylogenetic clade of

picocyanobacteria from the Mazurian lakes (Poland) reflects the early ontogeny of

glacial lakes." FEMS Microbiol Ecol. 2011 Jan; 75(1):89-98.

[11] Kahn K, Mazel D, Houmard J, Tandeau de Marsac N, Schaefer MR. "A role for cpeYZ in

cyanobacterial phycoerythrin biosynthesis." J Bacteriol. 1997 Feb; 179(4):998-1006.

[12] Liu JY, Jiang T, Zhang JP, Liang DC. "Crystal structure of allophycocyanin from red algae

Porphyra yezoensis at 2.2-A resolution." J Biol Chem. Jun 11, 1999; 274(24):16945-52.

[13] Murray JW, Maghlaoui K, Barber J. "The structure of allophycocyanin from

Thermosynechococcus elongatus at 3.5 A resolution." Acta Crystallogr Sect F Struct Biol

Cryst Commun. Dec 1, 2007; 63(Pt 12):998-1002.

[14] Padyana AK, Bhat VB, Madyastha KM, Rajashankar KR, Ramakumar S. "Crystal structure

of a light-harvesting protein C-phycocyanin from Spirulina platensis." Biochem Biophys

Res Commun. Apr 13, 2001; 282(4):893-8.

[15] Reuter W, Wiegand G, Huber R, Than ME. "Structural analysis at 2.2 A of orthorhombic

crystals presents the asymmetry of the allophycocyanin-linker complex, AP.LC7.8, from

phycobilisomes of Mastigocladus laminosus." Proc Natl Acad Sci U S A. Feb 16, 1999;

96(4):1363-8.

[16] Sidler W, Kumpf B, Suter F, Klotz AV, Glazer AN, Zuber H. "The complete amino-acid

sequence of the alpha and beta subunits of B-phycoerythrin from the rhodophytan alga

Porphyridium cruentum." Biol Chem Hoppe Seyler. 1989 Feb; 370(2):115-24.

[17] Stec B, Troxler RF, Teeter MM. "Crystal structure of C-phycocyanin from Cyanidium

caldarium provides a new perspective on phycobilisome assembly." Biophys J. 1999

Jun; 76(6):2912-21.

[18] Sugita C, Ogata K, Shikata M, Jikuya H, Takano J, Furumichi M, Kanehisa M, Omata T,

Sugiura M, Sugita M. "Complete nucleotide sequence of the freshwater unicellular

cyanobacterium Synechococcus elongatus PCC 6301 chromosome: gene content and

organization." Photosynth Res. 2007 Jul-Sep; 93(1-3):55-67.

[19] Troxler RF, Ehrhardt MM, Brown-Mason AS, Offner GD. "Primary structure of phycocyanin

from the unicellular rhodophyte Cyanidium caldarium. II. Complete amino acid sequence

of the beta subunit." J Biol Chem. Dec 10, 1981; 256(23):12176-84.

[20] Zhou J, Gasparich GE, Stirewalt VL, de Lorimier R, Bryant DA. "The cpcE and cpcF genes

of Synechococcus sp. PCC 7002. Construction and phenotypic characterization of

interposon mutants." J Biol Chem. Aug 15, 1992; 267(23):16138-45.

[21] de Lorimier R, Bryant DA, Stevens SE Jr. "Genetic analysis of a 9 kDa

phycocyanin-associated linker polypeptide." Biochim Biophys Acta. Aug 9, 1990;

1019(1):29-41.

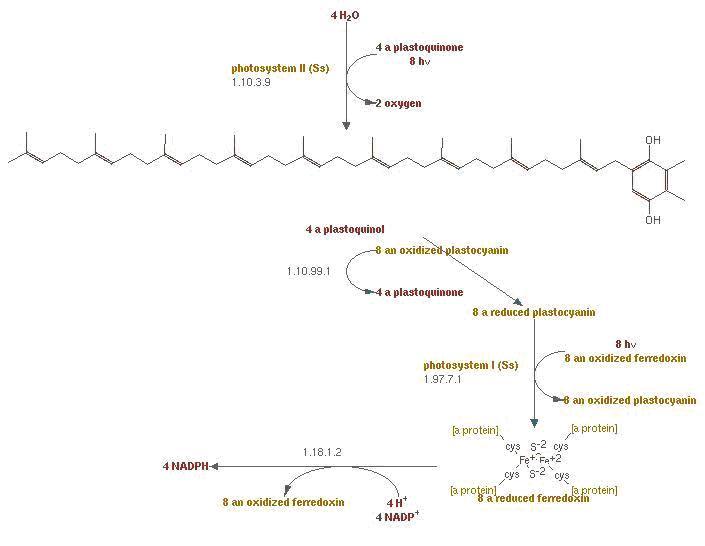
## Light Reactions

Photosynthesis is composed of two processes, the light reactions and the dark reactions. The light reactons take place in the two photosystems - photosystem I and photosystem II, where light energy is harvested and is used to power the transfer of electrons from water, via a series of electron donors and acceptors, to the final acceptor NADP+, which is reduced to NADPH. The NADPH generated by the light reactions is used for sugar synthesis in the dark reactions.

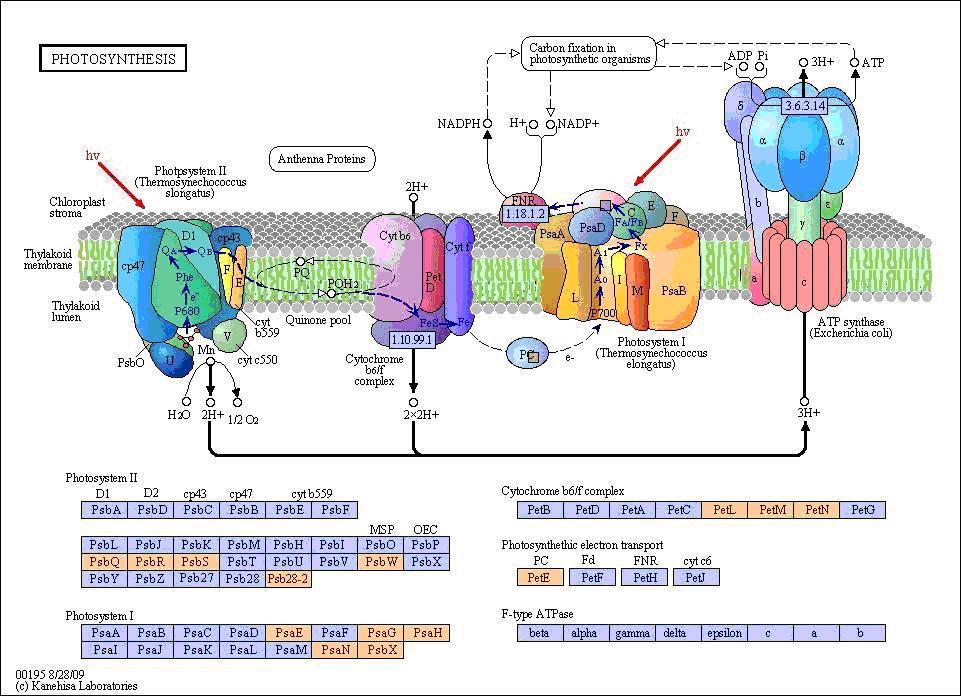
There are two general chemical reactions involved in the light reactions: water oxidation in photosystem II, and NADP reduction in photosystem I. Both of the photosystems are large multiprotein complexes contained within the thylakoid membranes of all types of plants, algae and cyanobacteria, and both photosystems capture the light by means of large antennae systems, consisting of chlorophylls and carotenoids.

Two of the EC numbers, 1.10.3.9 and 1.97.7.1, refer to photosystems II and I, respectively. Each of these photosystems consist of many parts, and each of the parts were checked to see if they existed in *Microcoleus vaginatus*.

**Automated Annotation**



S Fig. 1: MetaCyc illustration of the light reactions



S Fig. 2: Kegg S Figure showing the proteins predicted by IMG/ER to be in *Microcoleus vaginatus* (blue) and to be missing in *Microcoleus vaginatus* (orange)

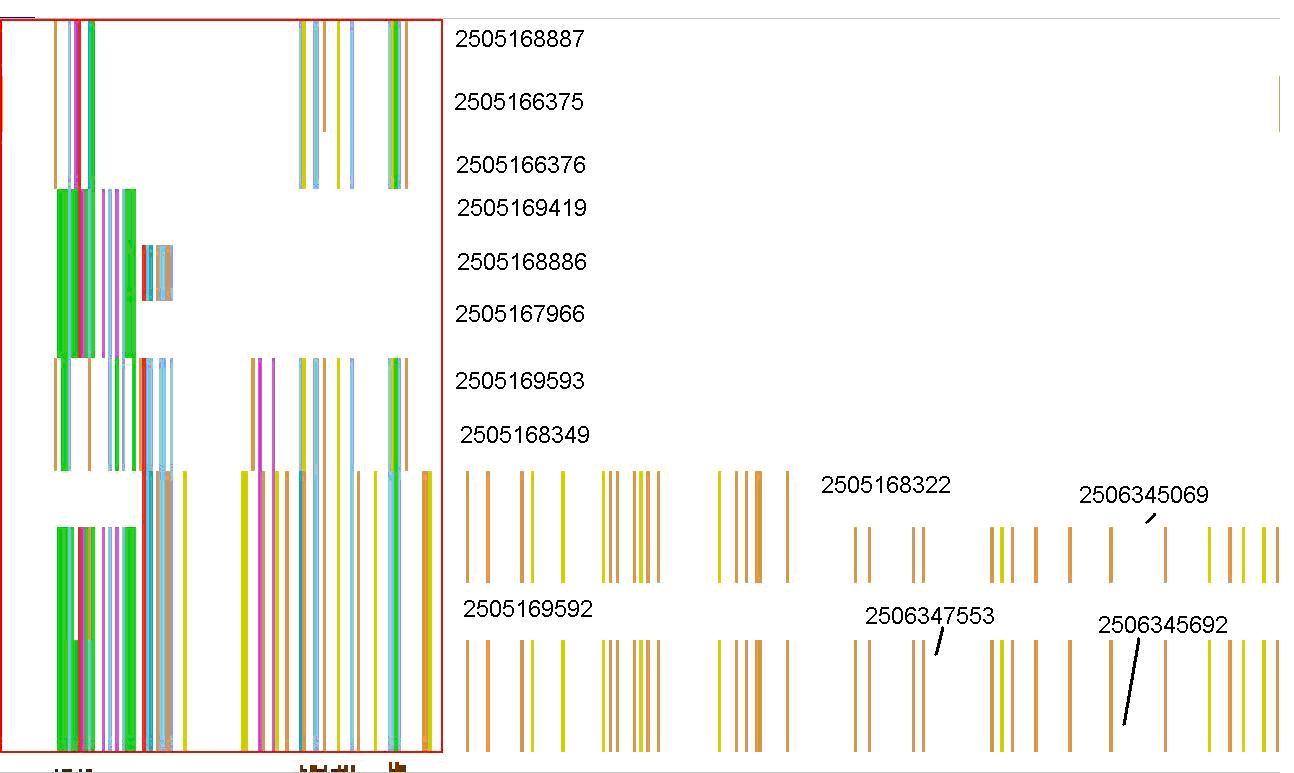
**PHOTOSYSTEM II EC: 1.10.3.9**

**PsbA**: **Photosystem II Reaction Center D1 protein**

Gene OIDs: 2505167966, 2505169419, 2505168887, 2505168886, 2505166376, 2505166375, 2505169593, 2505168349, 2505169592, 2505168322

No definitive prediction has been reached as to which genes in *M. vaginatus* code for psbA (D1). Predicted genes via KEGG and homologs to related cyanobacteria *Synechocystis sp.* PCC 6803 [1] and *M. vaginatus* FGP-2, although possessing strong alignment, only existed in fragments at scaffold endings (usually representing a bad read) and were much short of the typical 360 amino acid length. This contributes a sizable amount of uncertainty to further annotation of psbA.

Cyanobacteria are known to possess multiple versions of psbA [6, 7] that are expressed in different situations. Differing gene neighborhoods among the fragments supports this; up to 10 different versions of psbA may exist, but because some fragments are the sole occupant of their scaffold some fragments may belong to the same gene. A multiple sequence alignment (via MUSCLE) with the 10 PCC 9802 and 3 FGP-2 psbA genes indicates high conservation amongst different groups of fragments, suggesting different classes of D1. An active site check also remains difficult due to the fragmentation.



S Fig. 3: Overview of MUSCLE multiple sequence alignment. Colored bands represent matching amino acids with similar properties across one or more protein sequences.

**PsbB: Photosystem II chlorophyll-binding protein CP47**

Gene OID: 2505165716

Gene was predicted to code for the PsbB protein. A NCBI BLAST search [2] (NCBI: P20093) gave another PsbB proteins as a top result with an e-value of 0 and 78% identity. A PDB BLAST search [3] (PDB: 1S5L) gave an article about PS II, and it mentioned active site residues F362 and F363, which both matched with the gene in M. vaginatus. A reverse BLAST search gave gene 2505165716 as the top hit.

**PsbC: Photosystem II protein CP43**

Gene OIDs: 2505168232 and 2505170590

Gene 2505168232 was predicted to code for the PsbC protein. A NCBI BLAST search (NCBI: P09193) [4] gave other PsbC proteins as top results with e values of 0 and close to 100% query coverage. A PDB BLAST search (PDB: 1S5L) [3] gave an article about PSII, and it mentioned active site residues R357 and E354, which both matched with the gene in M. vaginatus. Additionally, a conserved Gly-Gly-Glu-Thr-Met-Arg-Phe-Trp-Asp motif somewhere in the gene was mentioned, and such a motif was found in the M. vaginatus gene. A reverse BLAST search gave gene 2505168232 as the top hit.

Gene 2505170590 was also confirmed to code for the PsbC protein. A NCBI BLAST search (NCBI: B1VKG9) [4] yielded other Photosystem II CP43 chlorophyll apoproteins as top results with an e-value of 3e-53 and high query coverage of 92%. A reverse BLAST search gave gene 2505168232 as the top hit and then gene 2505170590 as the second hit.

**PsbD: Photosystem II DII Subunit Protein**

Gene OID: 2505168220

The gene was predicted to code for the PsbD protein. A NCBI BLAST search (NCBI: Q8XFA5) [5] gave as top results other photosystem II D2 proteins with a low e-value of 5e-177 and a high query coverage of 100%. A reverse BLAST search against the *Microcoleus vaginatus* PCC 9802 genome gave gene as a top hit, with an e-value of 0.

**PsbE: Photosystem II Cytochrome b559, Alpha Subunit Protein**

Gene OID: 2505166666

The gene was predicted to code for the PsbE protein. A NCBI BLAST search (NCBI: B0JLV1) [6] gave as top results other photosystem II cyt b559 alpha subunit proteins with an e-value of 2e-33 and high query coverage of 96%. A reverse BLAST search against the *Microcoleus vaginatus* PCC 9802 genome gave gene 2505166666 as a top hit, with an e-value of 2e-35.

**PsbF: Photosystem II Cytochrome b559, Beta Subunit Protein**

Gene OID: 2505166665

The gene was predicted to code for the PsbF protein. A NCBI BLAST search (NCBI: Q8YQI1) [7] gave as top results other photosystem II cyt b559 beta subunit proteins with an e-value of 1e-13 and high query coverage of 100%. A reverse BLAST search against the *Microcoleus vaginatus* PCC 9802 genome gave gene as a top hit, with an e-value of 1e-15.

**PsbH: Photosystem II 10 kDa phosphoprotein**

Gene OID: 2505168296

The gene was predicted to code for PsbH protein. A NCBI BLAST search (NCBI: Q4G3C2) [5] gave as top results entries for photosystem II reaction center protein H, with an e-value of 7e-23 and high query coverage of 97%. A reverse BLAST search against the *Microcoleus vaginatus* genome gave gene 2505168296 as a top hit, with an e-value of 2e-16.

**PsbI: Photosystem II reaction centre I protein**

Gene OID: 2505165663

The gene was predicted to code for the PsbI protein. A NCBI BLAST search [] (NCBI: A0T0U9) [8] gave as top results entries for photosystem II reaction center protein I, with an e-value of 5e-10 and high query coverage of 82%. A reverse BLAST search against the *Microcoleus vaginatus* PCC 9802 genome gave gene 2505165663 as a top hit with an e-value of 2e-08.

**PsbJ: Photosystem II Reaction Center Protein J**

Gene OID: 2505166663

The gene was predicted to code for the PsbJ protein. A NCBI BLAST search (NCBI: P59087) [9] gave as top results entries for photosystem II reaction center protein J, with low e-values (around 1e-10) and high query coverage (above 90%). A reverse BLAST query also gave the gene as a top hit with an e-value of 7e-13.

**PsbK: Photosystem II Reaction Center Protein K**

Gene OID: 2505165902

The gene might code for the PsbK protein. A NCBI BLAST search [10] (NCBI: P14089) gave as top results entries for Photosystem II reaction center protein K, with low e-values (around 1e-14) and high query coverage (100%). However, the reverse BLAST query against the *Microcoleus vaginatus* PCC 9802 genome gave no hits.

**PsbL: Photosystem II PsbL Protein**

Gene OID: 2505166664

The gene was predicted to code for the PsbL protein. A NCBI BLAST search (NCBI: Q46H80) [11] gave as top results entries for photosystem II reaction center protein L, with low e-values (around 1e-10) and high query coverage (around 90%). A reverse BLAST query also gave the same gene as a top hit with an e-value of 4e-12.

**PsbM: Photosystem II reaction center protein PsbM**

Gene OID: 2505170079

The gene was predicted to code for the PsbM protein. A NCBI BLAST search [5] (NCBI: Q8YYG7) gave as top results entries for Photosystem II reaction center protein M, with an e-value of 1e-11 and high query coverage of 100%. A reverse BLAST search gave gene 2505170079 as the top hit.

**PsbO: Manganese-stabilising protein / photosystem II polypeptide**

Gene OID: 2505169490

The gene was predicted to code for the PsbO protein. A NCBI BLAST search [12] (NCBI: Q9R6W6) gave as top results other entries for photosystem II manganese-stabilizing polypeptides, with an e-value of 2e-95 and high query coverage of 98%. A reverse BLAST search also gave the gene as a top hit with an e-value of 9e-92.

**PsbP: Photosystem II OEC (Oxygen-Evolving enhancer protein, Chloroplastic)**

Gene OID: 2505166392

The gene was predicted to code for the PsbP protein. A NCBI BLAST search [13] (NCBI: P16059) gave as top results other entries for Oxygen-evolving enhancer protein 2, chloroplastic, with an e-value of 7e-08 and high query coverage of 80%. A reverse BLAST search gave gene 2505166392 as the top hit.

**PsbQ:**

No gene was predicted by IMG/ER to code for the PsbQ protein. To double-check, a PsbQ protein FASTA sequence from a cyanobacteria was obtained from UniProt, and a BLAST search done with this sequence gave gene 2505169772 as a hit, with an e-value of 3e-09 and 25% identity. A NCBI BLAST search with this hit gave bad results. None of the other hits were PsbQ proteins. However, the literature suggests that PsbQ is found in cyanobacteria [14, 15], so further studies should be done for this gene and for the PsbQ protein.

**PsbR:**

No gene was predicted by IMG/ER to code for the PsbR protein. To double-check, a reviewed PsbR protein FASTA sequence from mouse-ear cress was obtained from UniProt. A BLAST search done with this sequence gave no genes as hits, so it suggests that IMG/ER was correct in not predicting any genes to code for the PsbR protein. The literature further verifies that PsbR is not found in green algae [16].

**PsbS:**

No gene was predicted by IMG/ER to code for the PsbS protein. To double-check, a reviewed PsbS protein FASTA sequence from mouse-ear cress was obtained from UniProt. A BLAST search done with this sequence gave no genes as hits, so it suggests that IMG/ER was correct in not predicting any genes to code for the PsbS protein. Additionally, the literature supports the conclusion that cyanobacteria do not have PsbS proteins. Cyanobacteria have another protein, OCP (orange carotenoid-binding protein) which performs a similar function to PsbS and which later evolved in plants to the PsbS protein [17].

**PsbT: Photosystem II reaction center T protein**

Gene OID: 2505165717

The gene was predicted to code for the PsbT protein. An NCBI BLAST search [5]( (NCBI: Q8Z0F9) gave other photosystem II reaction center T proteins as top results, with an e-value of 2e-08 and high query coverage of 100%. A reverse BLAST search against the *Microcoleus vaginatus* PCC 9802 genome gave gene 2505165717 as the top hit, with an e-value of 1e-09.

**PsbU: Photosystem II 12 kDa extrinsic protein**

Gene OID: 2505166404

The gene was predicted to code for the PsbU protein. An NCBI BLAST search [5](NCBI: Q8YXJ7) gave other photosystem II 12 kDa extrinsic proteins as top results, with an e-value of 9e-42 and high query coverage of 99%. A reverse BLAST search against the *Microcoleus vaginatus* PCC 9802 genome gave gene 2505166404 as the top hit, with an e-value of 7e-37.

**PsbV: Photosystem II Cytochrome PsbV**

Gene OID: 2505166307 and 2505166308

Both genes 2505166307 and 2505166308 were predicted to code for the PsbV protein. The NCBI BLAST search for gene [18] (NCBI: Q9ET95) gave a cytochrome c550 as the top result, with an e-value of 6e-42 and 80% query coverage. A reverse BLAST search gave gene 2505166307 as the top hit, and gene 2505166308 as the second hit. The NCBI BLAST search for gene [19] (NCBI: B0C731) also gave a cytochrome c550 hit, with an e-value of 4e-53 and 100% query coverage. A reverse BLAST search gave gene 2505166308 as the top hit, and gene 2505166307 as the second hit.

**PsbW:**

No gene was predicted by IMG/ER to code for the PsbW protein. To double-check, a reviewed cyanobacteria PsbW protein FASTA sequence was obtained from UniProt [UNI]. A BLAST search done with this sequence gave gene 2505169335 as a hit, with an e-value of 3e-38 and 62% identity. This gene was actually also a hit for protein Psb28, and as mentioned for protein Psb28, all the NCBI BLAST hits were for Psb28 proteins. However, since the e-value for this hit to the gene is very good, it is possible that the Psb28 protein is bifunctional for both Psb28 and PsbW. The literature does suggest however, that PsbW is not found in cyanobacteria [20] but is found in plants, so this would definitely be a candidate for further research.

**PsbX: Photosystem II reaction center X protein**

Gene OID: 2505166793

The gene was predicted to code for the PsbX protein. An NCBI BLAST search [5] (NCBI: Q8YYB0) gave other photosystem II reaction center X proteins as top results, with an e-value of 2e-10 and high query coverage of 100%. A reverse BLAST search gave gene 2505166793 as the top hit.

**PsbY: Photosystem II protein Y (PsbY)**

Gene OID: 2505165632

The gene might code for the PsbY protein. A NCBI BLAST search [12] (NCBI: B1WXD2) gave as top results other entries for Photosystem II protein Y, with an e-value of 9e-07 and high query coverage of 92%. A reverse BLAST search gave no hits.

**PsbZ: Photosystem II core protein PsbZ**

Gene OID: 2505166154

The gene was predicted to code for the PsbZ protein. A NCBI BLAST search [6] (NCBI: B0JLL9) gave as top results other entries for Photosystem II reaction center protein Z, with an e-value of 2e-17 and high query coverage of 98%. A reverse BLAST search also gave the gene as a top hit, with an e-value of 1e-08.

**Psb27: Photosystem II protein Psb27**

Gene OID: 2505169792

The gene was predicted to code for the Psb27 protein. A NCBI BLAST search (NCBI: P74367) [4] gave as a top result a Photosystem II 11 kDa protein, which is another name for the Psb27 protein. This hit had a very low e-value (8e-34) and good query coverage (99%). A reverse BLAST search also gave the gene as a top hit, with an e-value of 3e-28.

**Psb28: Photosystem II reaction center protein Psb28**

Gene OID: 2505169335

The gene was predicted to code for the Psb28 protein. A NCBI BLAST search [5] (NCBI: Q8YYP5) gave as top results other photosystem II reaction center Psb28 proteins, with an e-value of 3e-44 and high query coverage of 100%. A reverse BLAST query also gave the gene as the first hit.

**Psb28-2:**

No gene was predicted by IMG/ER to code for the Psb28-2 protein. To double-check, a cyanobacteria Psb28-2 FASTA seqeunce was obtained from UniProt, and a BLAST search was done against the *Microcoleus vaginatus* genome. The search gave as a hit gene 2505169335 with an e-value of 4e-09 and 31% identity. This gene is the same gene as for protein Psb28, and is also the gene that was predicted to possibly code for PsbW. It is less likely that the gene codes for Psb28-2 due to the lower e-value and percent identity, but it is still possible. Lab work needs to be done to see if Psb28-2 does exist in *Microcoleus vaginatus*, and to see if gene 2505169335 encodes for it or another gene does.

**PHOTOSYSTEM I: EC 1.97.7.1**

**PsaA: Photosystem I P700 chlorophyll a apoprotein A**

Gene OID: 2505169847

Gene 2505169847 is predicted to code for psaA due to its strong homology with other psaA proteins found in cyanobacteria in PDB (1JBO) [21] and NCBI SwissProt (P29254) [22] databases. The top hits in PDB and NCBI both had e-values of 0. The PDB hit had 83% percent identity while the NCBI hit had 100% query coverage. A reverse BLAST search gave gene 2505169847 as the top hit.

**PsaB: Photosystem I P700 chlorophyll a apoprotein A**

Gene 2505169846 is predicted to code for psaB due to its strong homology with other psaB proteins found in cyanobacteria in PDB (1JBO) [21]and NCBI SwissProt (P58565) [5] databases. The top hits in PDB and NCBI both had e-values of 0. The PDB hit had 87% identity while the NCBI hit hid 100% query coverage. A revese BLAST search gave gene 2505169846 as the top hit.

**PsaC: Photosystem I iron-sulfer center**

Gene 2505168274 is predicted to code for psaC due to its strong homology with verified psaC genes found in cyanobacteria in PDB (1K0T) [23] and SwissProt databases (P0A412.2) [24]. The e-values were low and percent identity were high. For the NCBI and PDB hits, the e-values were 1e-40 and 3.49159e-40 respectively. The PDB hit had 96% identity, while the NCBI hit had 100% query coverage. A reverse BLAST search gave gene 2505168274 as the top hit.

**PsaD: Photosystem I reaction centre subunit II**

Gene OID: 2505171069

The gene was predicted to code for the PsaD protein. A NCBI BLAST search (NCBI: P0A420) [25] gave as a top result a photosystem I reaction center subunit II, with 97% query coverage and an e-value of 1e-53. A reverse BLAST query also gave the gene as the first hit.

**PsaE: Photosystem I reaction centre subunit IV**

Gene OID: 2505168753

There was no gene predicted by IMG/ER to code for the PsaE protein. To double-check, a reviewed cyanobacteria FASTA sequence was taken from UniProt, and a BLAST search was done against the *Microcoleus vaginatus* genome in IMG/ER. The BLAST search gave gene 2505168753 as a hit with an e-value of 4e-16 and 62% identity. A NCBI BLAST search (NCBI: Q9WWP1) [26] was then done with this gene, and the search gave as a top result a Photosystem I reaction centre subunit IV with an e-value of 1e-25 and 54% query coverage. This hit isn’t as strong as the other hits, but it is still probable that the PsaE protein does exist in *Microcoleus vaginatus*. A reverse BLAST search gives gene 2505168753 as the top hit.

**PsbF: Photosystem I reaction centre subunit III**

Gene OID: 2505165725

The gene was predicted to code for the PsaF protein. A NCBI BLAST search (NCBI: P31091) [27] gave as a top result a Photosystem I reaction center subunit III, with 100% query coverage and an e-value of 4e-58. A reverse BLAST query also gave the gene as the first hit.

**PsaG:**

There was no gene predicted by IMG/ER to code for the PsaG protein. To double-check, a reviewed cyanobacteria FASTA sequence was taken from IMG/ER and a BLAST search was done against the *Microcoleus vaginatus* genome in IMG/ER. The BLAST search gave no genes as hits, so it seems that IMG/ER was correct in predicting no genes to code for the PsaG protein. A literature search further revealed that protein PsaG was only found in higher order seed plants [28].

**PsaH:**

There was no gene predicted by IMG/ER to code for the PsaH protein. To double-check, a reviewed cyanobacteria FASTA sequence was taken from IMG/ER and a BLAST search was done against the *Microcoleus vaginatus* genome in IMG/ER. The BLAST search gave no genes as hits, so it seems that IMG/ER was correct in predicting no genes to code for the PsaH protein. A literature search further revealed that protein PsaH was only found in higher order seed plants [28].

**PsaI: Photosystem I reaction center subunit VIII**

Gene OID: 2505168149

The gene was predicted to code for the PsaI protein. A NCBI BLAST search (NCBI: P0A427) [21] gave as a top result a Photosystem I reaction center subunit VIII, with 100% query coverage and an e-value of 6e-10. A reverse BLAST query also gave the gene as the first hit.

**PsaJ: Photosystem I reaction center subunit IX**

Gene OID: 2505165726

The gene is predicted to code for the Photosystem I reaction center subunit IX/ PsaJ protein. An NCBI BLAST search (NCBI: P58568) [5] gave other Photosystem I reaction center subunit IX/ PsaJ proteins as top results. The top hit had an e-value of 2e-10 and high query coverage of 97%. A reverse BLAST search against the *Microcoleus vaginatus* PCC 9802 genome gave gene 2505165726 as a top hit, with an e-value of 2e-12.

**PsaK: Photosystem 1 reaction center subunit**

Gene OID: 2505167587 and 2505170413

The gene is predicted to code for the Photosystem I reaction center subunit PsaK protein. An NCBI BLAST search (NCBI: P0A425) [21] gave other Photosystem I reaction center subunit PsaK proteins as top results. The top hit had an e-value of 1e-21 and high query coverage of 90%. A reverse BLAST search against the *Microcoleus vaginatus* PCC 9802 genome gave gene 2505167587 as a top hit, with an e-value of 1e-23.

IMG/ER predicts gene 2505170413 to code for a hypothetical protein. No clear conclusion can be made about gene 2505170413, but it might code for a subunit of the PsaK protein, or it might also, like gene 2505167587, code for the Photosystem I reaction center subunit PsaK protein. An NCBI BLAST search (NCBI: Q8YLK8) [5] gave other Photosystem I reaction center subunit PsaK proteins as top results. The top hit had an e-value of 3e-15 and query coverage of 58%. A reverse BLAST search against the *Microcoleus vaginatus* PCC 9802 genome gave gene 2505170413 as a top hit, with an e-value of 5e-15.

**PsaL: Photosystem I reaction center subunit XI**

Gene OID: 2505168150

The gene 2505168150 is predicted to code for the Photosystem I reaction centre subunit XI protein. An NCBI BLAST search (NCBI: Q8DGB4) [21] gave other Photosystem I reaction centre subunit XI proteins as top results. The top hit had a low e-value of 2e-48 and high query coverage of 95%. A reverse BLAST search against the *Microcoleus vaginatus* PCC 9802 genome gave gene 2505168150 as a top hit, with an e-value of 2e-45.

**PsaM:**

Gene OID: 2505168179

No clear conclusion can be made about gene 2505168179, but it might code for a subunit of the PsaM protein. An NCBI BLAST search (NCBI: P0A403) [21] gave other Photosystem I reaction center subunit PsaM proteins as top results. The top hit had a high e-value of 1e-7 but a high query coverage of 100%. However, a reverse BLAST search against the *Microcoleus vaginatus* PCC 9802 genome gave no hits. Further research is needed to determine whether or not this gene (2505168179) actually codes for the Photosystem I reaction center subunit PsaM protein.

**PsaN:**

No clear conclusion can be made about which gene codes for the Photosystem I reaction center subunit PsaM protein, but it might code for a subunit of the PsaN protein. IMG/ER did not predict any genes within *Microcoleus vaginatus* PCC 9802 to code for this protein. This protein was searched on UniProt but no cyanobacteria within UniProt were found to have the Photosystem I reaction center subunit PsaM protein. It is very likely that the gene to code for this protein simply does not exist within *Microcoleus vaginatus.* A literature search further revealed that protein PsaN was only found in higher order seed plants 18713342 [28].

**PsaX:**

Gene OID: 2505168727

No clear conclusion can be made about gene 2505168727, but it might code for a subunit of the PsaX protein. Initially IMG/ER did not predict any genes to code for the Photosystem I reaction center subunit PsaX protein. The gene coding for this protein was found using Uniprot. This gene was then BLAST-ed against the *Microcoleus vaginatus* PCC 9802 genome and the gene 2505168727 was found. An NCBI BLAST search (NCBI: P58566) [5] gave other Photosystem I 4.8 kDa proteins as top results. The top hit had a high e-value of 7e-7, but a high query coverage of 85%. However, a reverse BLAST search against the *Microcoleus vaginatus* PCC 9802 genome gave no hits. Further research is needed to determine whether or not this gene (2505168727) actually codes for the Photosystem I reaction center subunit PsaX protein, although the literature suggests that PsaX should exist in *Microcoleus vaginatus*, since PsaX is specific for thermophilic cyanobacteria [29].

**PHOTOSYNTHETIC ELECTRON TRANSPORT**

***Cytochrome b6f complex EC 1.10.99.1***

**PetA: Apocytochrome F**

Gene OID: 2505169770

The gene is predicted to code for the apocytochrome F subunit of cytochrome b6f. An NCBI BLAST (NCBI: P13626) [30] reveals extremely strong homology to *Nostoc* (e-value of 2e-145 and 100% query coverage). No alignment data exists on PDB. A reverse BLAST search with the NCBI BLAST sequence against the *Microcoleus vaginatus* genome gave gene 2505169770 as the top hit.

**PetB: Cytochrome b subunit of the bc complex**

Gene OID: 2505168501

The gene is predicted to code for the “b” subunit of cytochrome b6f. NCBI BLAST reveals very strong homology with *Cyanothece* sp., with an e-value of 2e-116 and 100% query coverage (NCBI: B1WWK9) [12]. No alignment data exists on PDB. A reverse BLAST search gave gene 2505168501 as the top hit.

**PetC: Cytochrome b6-f complex iron-sulfur subunit**

Gene OID: 2505169769

This gene is predicted to code for the iron-sulfur cytochrome subunit. A NCBI BLAST search gave as a top hit another iron-sulfur cytochrome subunit as the top hit, with an e-value of 4e-82 and 100% query coverage (NCBI: B8HNR1) [31]. A reverse BLAST search gave gene 2505169769 as the top hit. Also, a PDB BLAST search (PDB: 2E74 [32]) gave an article about an iron-sulfur cytochrome subunit. The active site residues mentioned in this article matched up with the ones in *Microcoleus vaginatus*.

**PetD: Cytochrome b6-f complex subunit IV**

Gene OID: 2505168502

This gene is predicted to code for subunit IV of cytochrome b6f. This is supported by very strong homology with the corresponding protein subunit in *Microcystis* as determined by an NCBI BLAST (NCBI: B0JM93) [6], with an e-value of 3e-76 and a query coverage of 100%, as well as strong homology with *Nostoc*, as determined by a PDB BLAST (PDB: 2ZT9), with an e-value of 1.91446E-74 and a percent identity of 84%. A reverse BLAST search with the NCBI sequence gives gene 2505168502 as the top hit.

**PetG: Cytochrome B6-F complex subunit 5**

Gene OID: 2505170453

The gene was predicted to code for cytochrome b6f complex subunit 5. A PDB BLAST (PDB: 2ZT9) showed a corresponding protein in the related cyanobacterium *Nostoc* PCC7120, with a good alignment (e-value: 1.11e-10 with 83% identities). An NCBI BLAST (NCBI: Q113E9) [33] revealed a similar match (e-value: 8e-11 and 89% query coverage). The sequence of this subunit is only 39 aa long, so the high rate of identities is not surprising. A reverse BLAST with the NCBI sequence gives no genes as hits.

**PetL - Not existent in *Microcoleus***

In *Synechococcus*, PetL is dispensable and no gene codes for it. Given the taxonomic similarity between *Microcoleus vaginatus* and *Synechococcus*, in conjunction with the fact that the automated annotation did not find a gene corresponding to PetL in *Microcoleus vaginatus*, it is safe to assume that this gene does not exist in *Microcoleus vaginatus*. Furthermore, a paper on *Synechocystis PCC 6803* revealed that subunit PetL is dispensable and not necessary for the cytochrome to function [34].

**PetM:**

IMG did not predict a gene for PetM in *Microcoleus*. However, BLASTing a PetM from *Synechocystis* (gene OID: 637012267) against the *M. vaginatus* PCC9802 genome revealed an interesting “hypothetical protein” (gene OID: 2505168468). Running a multiple sequence alignment reveals that there are some highly conserved sections between the two genes, but the gene in *M. vaginatus* has an extremely long insertion in between two of the conserved sections, drastically increasing the e-value of any BLAST. The *Synechocystis* gene is only 36aa long, whereas the “hypothetical protein” in *M. vaginatus* is 430aa long. This part of the genome should be sequenced again and rechecked, since it’s possible that sequencing errors caused the errors here. However, a literature search showed that the PetM subunit is not essential for the cytochrome to function [35].

**PetN:**

Similar to the results for PetM, IMG could not predict a gene for PetN. Blasting a PetN sequence from *Synechococcus elongatus* (gene OID: 637616704) versus *M. vaginatus* reveals a gene which, like the one discussed above in PetM, shares many highly conserved sections with the known PetN sequence, but includes very large insertions (gene OID: 2505169403). This part of the genome should be sequenced again and rechecked, since it’s possible that sequencing errors caused the errors here. A literature search showed however, that PetN is a subunit critical for the cytochrome to function [34], so further research should be done to see if the PetN subunit does exist and whether or not another gene codes for it.

***Photosynthetic electron transport***

**PetE: Plastocyanin**

Just like PetN and PetM, IMG could not predict a gene for PetE, with exactly the same problems (namely, conserved sections separated by enormous insertions). An IMG genome blast of *Synechocystis* plastocyanin (gene OID: 637011725), as well as *Synechococcus* platocyanin (gene OID: 637799514) against the *Microcoleus* genome turned up a number of hits, none of which led to any conclusions. According to the literature, plastocyanin is functionally equivalent to the cytochrome c6 complex [36], so it is possible that plastocyanin is not found in *Microcoleus vaginatus*. This should be left for further investigation.

**PetF: Ferredoxin**

Gene OID: 2505166364, 2505166512, 2505167674 and 2505169638

Four genes were predicted by IMG/ER to code for the PetF protein. For gene 2505166364, a NCBI BLAST search (NCBI: P00245) [37] gave a ferredoxin as a top result, with an e-value of 7e-47 and 100% query coverage. For gene 2505166512, a NCBI BLAST search (NCBI: P08451) [38] gave a ferredoxin-2 as a top result with an e-value of 6e-24 and 75% query coverage. For gene 2505167674, a NCBI BLAST search (NCBI: P15788) [39] gave a ferredoxin as a top result, with an e-value of 5e-31 and 93% query coverage. For gene 2505169638, a NCBI BLAST search (NCBI: P08451) [38] gave a ferredoxin-2 as a top result, with an e-value of 6e-30 and 96% query coverage. Reverse BLAST searches were also done. The NCBI hit for gene 2505166364 gave as a top result gene 2505166364. The NCBI hit for gene 2505166512 gave as a top result gene 2505169638, and then gene 2505166512. The NCBI hit for gene 2505167674 gave as a top result gene 2505166364, and then gene 2505167674. The NCBI hit for gene 2505169638 gave as a top result gene 2505169638. The four genes could all code for the same ferredoxin, or they could each code for parts of the ferredoxin. An experiment would need to be done to test this and see if the genes all code for the same ferredoxin or if they code for parts of the whole. From the reverse BLAST searches, it would seem that 2505166364 is most likely to code for a ferredoxin, followed by gene 2505169638. We are not as sure about the other genes, but all the genes could potentially code for ferredoxin.

**PetH: Ferredoxin--NADP(+) reductase**

Gene OID: 2505171062 EC: 1.18.1.2

Gene was predicted to code for ferredoxin NADP reductase. A NCBI BLAST search gave other ferredoxin NADP reductases as results with an e-value of 3e-155 and a query coverage of 97% (NCBI: Q93RE3) [40]. A PDB BLAST search gave an article on the structure of ferredoxin NADP reductase with an e-value of 5.41678E-127 and a percent identity of 71%(PDB: 1GJR) [41]. All the active site residues matched, except for at site 233, where R was replaced by K in microcoleus vaginatus. R and K are similar amino acids, and according to the article, the purpose of R is to provide a side chain that can hydrogen bond. K can do this too, since it also has a nitrogen group on its side chain. Finally, a multiple sequence alignment was done against other cyanobacteria, and although most of them had a R at site 233, some did have a K at site 233.

**PetJ: Cytochrome C6**

Gene OID: 2505166306 and 2505166703

Both genes were predicted to code for protein PetF. For gene 2505166306, a NCBI BLAST search (NCBI: Q3MDW2) [42] gave as a top result a Cytochrome c6, with an e-value of 3e-37 and query coverage of 99%. For gene 2505166703, a NCBI BLAST search (NCBI: Q8Z0D7) [5] gave as a top result a cytochrome c6, with an e-value of 1e-29 and 96% query coverage. Reverse BLAST searches gave the respective genes as first hits, with the other gene as a second hit. There are 2 genes that potentially code for this same cytochrome. This might be simply because there are redundant genes, or because the genes code for different subunits of the cytochrome.

**F-type ATP Synthase EC: 3.6.3.14**

**Alpha subunit**

Gene OID: 2505169630

The gene is predicted to code for the ATPase alpha subunit due to its strong homology with verified alpha subunit genes found in cyanobacteria in PDB (1FXO) [43] (e-value of 0 and 75% identity) and Swissprot databases (Q05372.4) [44] (e-value of 0 and 99% query coverage). A reverse BLAST search with the Swissprot sequence gave gene 2505169630 as the top result.

**Beta Subunit**

Gene OID: 2505169408

The gene is predicted to code for the ATPase beta subunit due to its strong homology with verified gamma subunit genes found in cyanobacteria in PDB (1FXO) [43] (e-value of 0 and 80% identity) and Swissprot databases (P06540.2) [45] (e-value of 0 and 99% query coverage). A reverse BLAST search with the Swissprot sequence gave gene 2505169408 as the top hit.

**Gamma subunit**

Gene OID: 2505169631

The gene is predicted to code for the ATPase gamma subunit due to its strong homology with verified gamma subunit genes found in cyanobacteria in PDB (2QE7) [46] (e-value of 2.65132E-45 and 34% identity) and Swissprot databases (P12408.2) [2900236] (e-value of 4e-148 and 99% query coverage). A reverse BLAST search with the Swissprot sequence gave gene 2505169631 as the top hit. Out of two suspected sites (Arg10, Arg9) crucial to this subunit, only site 10 matched. Site 9 had D9 substituted for R9.

**Delta subunit**

Gene OID: 2505169629

The gene is predicted to code for the ATPase delta subunit due to its strong homology with verified Delta subunit genes found in cyanobacteria in Swissprot databases (P12406.1) [47] (e-value of 9e-55 and 97% query coverage). A reverse BLAST search gave gene 2505169629 as the top hit. No homologous structures were found in PDB most likely due to the difficulty of crystallizing membrane proteins.

**Εpsilon subunit**

Gene OID: 2505169407

The gene is predicted to code for the ATPase epsilon subunit due to its strong homology with verified Epsilon subunit genes found in cyanobacteria in PDB (2RQ6) [48] (e-value of 5.91516E-37 and 80% identity) and Swissprot databases (P26533.3) [49] (e-value of 3e-45 and 64% identity). A reverse BLAST search with the Swissprot sequence gave gene 2505169407 as the top hit.

**“A” Subunit**

Gene OID: 2505169625

The gene is predicted to code for the ATPase A subunit due to its strong homology with verified A subunit genes found in cyanobacteria in Swissprot databases with an e-value of 2e-100 and a query coverage of 95% (P12404.2) [47]. No strong homologs were found in PDB, once again most likely due to the difficulty of crystallizing membrane proteins. A reverse BLAST search with the Swissprot sequence gave gene 2505169625 as the top hit.

**“B” Subunit**

Gene OID: 2505169627 and 2505169628

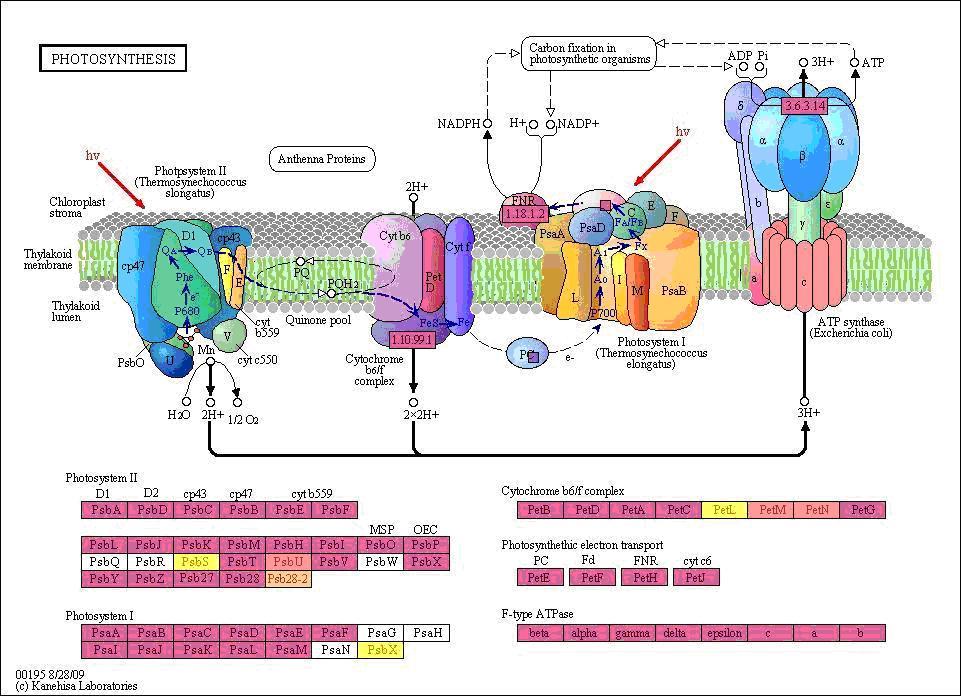
Two genes, 2505169627 and 2505169628, are predicted to code for ATPase B subunit (F1F0 and F0) due to the strong homology they shared with verified B subunits in Swissprot (P12410.1, Q05365.1) [44, 47] respectively. For gene 2505169627, the NCBI hit had an e-value of 2e-51 and 100% query coverage. For gene 2505169628, the NCBI hit had an e-value of 8e-42 and 100% query coverage. Reverse BLAST searches with the Swissprot sequences gave the respective genes as the top hits.

**“C” Subunit**

Gene OID: 2505169626

Gene 2505169626 is predicted to code for the ATPase A subunit due to its strong homology with verified A subunit genes found in cyanobacteria in PDB (2WIE) [50], with an e-value of 7.94842E-23 and a percent identity of 98%, and Swissprot (P08212.1) [51], with an e-value of 1e-32 and a query coverage of 98%. A reverse BLAST search with the Swissprot sequence gave gene 2505169626 as the top hit.

**Comparison with other Genomes**



S Fig. 4: Kegg S Figure showing IMG/ER predictions of proteins found and not found in cyanobacteria

Comparing the *Microcoleus vaginatus* genome to other cyanobacteria genomes in IMG/ER, there were some proteins predicted to be in other cyanobacteria that were not predicted to be in *Microcoleus vaginatus.* For photosystem II, protein PsbS was predicted to be in a small amount of cyanobacteria. It seems that *Microcoleus vaginatus* is part of the majority of cyanobacteria that don’t have the protein. Protein Psb28-2 was predicted to be in a small amount of cyanobacteria as well. *Microcoleus vaginatus* may or may not have a gene coding for the Psb28-2 protein, as mentioned earlier. Lab experiments would need to be done further verify or refute this claim.

For photosystem I, protein PsaE was not predicted to be in *Microcoleus vaginatus*, but it is predicted to be in most cyanobacteria. This is consistent with the PsaE protein that was found to be in *Microcoleus vaginatus* upon further searching. Protein PsaX was predicted to be in a small number of cyanobacteria, but was not predicted to be in *Microcoleus vaginatus*. Upon further searching, there was a hit for PsaX in the *Microcoleus vaginatus* genome, but this hit is uncertain. An experiment would need to be done to see if PsaX is expressed in *Microcoleus vaginatus*, and whether the gene codes for it or if another gene codes for PsaX.

For the cytochrome b6 complex, proteins PetL, PetM and PetN were not predicted to be in *Microcoleus vaginatus*. PetL was not found upon further search, while hypothetical proteins were found that might be proteins PetM and PetN. These proteins were found in some other cyanobacteria. Further research should be done to see if these proteins are essential for photosynthesis, and why *Microcoleus vaginatus* is missing these proteins while other cyanobacteria have them.

For photosynthetic electron transport, protein PetE was not predicted to be found nor was it found to be in *Microcoleus vaginatus*, while it is found in a majority of cyanobacteria. This warrants further research to explain why *Microcoleus vaginatus* is missing this protein while so many other cyanobacteria have it.

**Functional Studies**

Functional studies can be done on genes 2505168232 and 2505170590, which were predicted to code for protein PsbC. These genes should be studied to see if they both code for PsbC, or if each of them code for parts of PsbC. Two experiments should be done, with one of the genes being repressed in each experiment. It should then be seen from the experiments if the protein PsbC is still expressed when one or the other gene is repressed.

Functional studies should also be done on gene 2505169335. This gene was predicted to code for protein Psb28 primarily, as well as possibly proteins PsbW and Psb28-2. An experiment needs to first be done to see if *Microcoleus vaginatus* expresses proteins PsbW and Psb28-2 or not. If they do, another experiment should be done where gene 2505169335 is repressed, and it should be observed if proteins PsbW or Psb28-2 are expressed or not. If either one of the proteins are not expressed, it means the gene codes for that protein. Otherwise, it means that another gene codes for the protein.

Functional studies should be done on genes 2505166307 and 2505166308. These genes were both predicted to code for protein PsbV. These genes should be studied to see if they are redundant and both code for PsbV, or if each of them code for unique parts of the PsbV protein. As for protein PsbC, two experiments should be done, with one of the genes being repressed in each experiment. It should then be seen from the experiments if the protein PsbV is still expressed when one or the other gene is repressed.

Functional studies should be done for gene 2505168753. This gene codes for the PsaE protein. There was no gene predicted by IMG/ER to code for this protein, so this gene should definitely be checked just for that reason. Additionally, the NCBI hits confirming the identity of the gene were not as strong as other hits for other proteins. To test whether or not this gene codes for the PsaE protein, an experiment should first be done to see if *Microcoleus vaginatus* expresses the PsaE protein. If it does, an additionally experiment should be done to see if PsaE is still expressed if the gene is repressed. If it is no longer expressed, then the gene codes for PsaE.

Functional studies should be done for gene 2505170413. This gene codes for a hypothetical protein that was assigned by IMG/ER to have PsaK protein function. An experiment should be done to see if this gene is essentially for the expression for the PsaK protein. The gene should be repressed, and it should be observed whether or not the PsaK protein is still expressed. If it is, that means the gene is not essentially for the expression of PsaK. If PsaK is not expressed, then the gene is essentially for its expression.

Functional studies should also be done for gene 2505168179. This gene is predicted to code for the PsaM protein. An experiment should be done where the gene is repressed, and it should be observed whether or not the protein is expressed still. If it is not, the gene codes for the PsaM protein.

Functional studies should also be done for gene 2505168727. This gene might code for the PsaX protein, which was not predicted to be present in *Microcoleus vaginatus*. First, an experiment should be done to see if *Microcoleus vaginatus* actually expresses the PsaX protein. If it does, another experiment should be done with the gene repressed, and if this causes the protein to be repressed as well, then the gene codes for the protein.

Functional studies should be done on gene 2505168468. This gene codes for a hypothetical protein, which might be the PetM protein. An experiment should first be done to see if the PetM protein is expressed for *Microcoleus vaginatus*. If it is, another experiment should be done with gene repressed. If the gene is repressed and no PetM protein is expressed, then the gene codes for the PetM protein.

Functional studies should also be done on gene 2505169403, a hypothetical protein that might be the PetN protein. An experiment should first be done to see if the PetN protein is expressed in *Microcoleus vaginatus*. If it is, another experiment should be done with the gene repressed, and if the protein is repressed as well, this means the gene codes for the PetM protein.

A functional study should be done to see if a plastocyanin is expressed in *Microcoleus vaginatus*. This seems like an important component of the electron transport chain, and it would be interesting if it turns out that *Microcoleus vaginatus* actually turns out to be missing it. If a plastocyanin is found however, further studies would need to be done to try and find the gene or genes that code for it.

Functional studies should be done on genes 2505166364, 2505166512, 2505167674 and 2505169638. These genes were all predicted to code for ferredoxin. Experiments should be done with different genes repressed to see if the ferredoxin is still expressed, and in this way it can be determined whether or not the different genes code for the ferredoxin.

Studies of the fragments aligning with psbA are recommended due to the unique resistance *M. vaginatus’* D1 protein might offer against photoinhibition [8] [9], enabled by a mutation reducing the redox potential of Q(b) and Q(a) sites and allowing nonradiative charge recombination during photosynthesis [10]. Further lab work clarifying the fragment sequences at the ends of scaffolds, discerning which genes encode for the different D1 proteins, and identifying which mutations are responsible for photoinhibition resistance would be of help in understanding *M. vaginatus*.

Finally, functional studies should be done for genes 2505166306 and 2505166703. These genes were both predicted to code for cytochrome c6. It is unclear whether or not they both code for the cytochrome or if they code for different parts of the cytochrome. Experiments should be done with one or the other gene repressed to see if the cytochrome is still expressed, and in this way the genes can be proved or disproved to code for the cytochrome.

**Discussion**

*Microcoleus vaginatus* was found to be missing many proteins predicted to be necessary for photosynthesis. In photosystem II, proteins PsbQ, PsbR, PsbS, PsbW and Psb28-2 were found to missing. Protein PsbQ was found in cyanobacteria and needed for photosynthesis 20210304, so experiments should be done to try and find the protein in *Microcoleus vaginatus*. Proteins PsbR and PsbS were found to not be in cyanobacteria 17367749 18021070, and their absence is documented in literature. Protein PsbW was not found in cyanobacteria but only higher order plants 21487931, so the gene that gave a hit for PsbW in *Microcoleus vaginatus* might have evolved in plants to code for the PsbW protein.

In photosystem I, proteins PsaG, PsaH and PsaN were found to be missing, but this is well-documented in the literature for cyanobacteria 18713342. Protein PsaX was also found to be missing, but probably should not be missing. PsaX is associated with thermophilic cyanobacteria 19007746, so it would make sense for *Microcoleus vaginatus* to have this protein. Further experiments should be done to see if the organism has PsaX protein or not.

For the cytochrome b6 complex, proteins PetL, PetM and PetN were found to be missing. PetL and PetM are not essential for the function of the cytochrome b6 complex 17224258 11278512, so it is not problematic that those proteins are missing. PetN, however, is essential for the cytochrome to function 17224258, so further experiments should be done to see if the protein is just missing because it is hard to find the gene coding for it, or if the protein is actually missing.

Finally, protein PetE, plastocyanin, was found to be missing. Plastocyanin is functionally equivalent to the cytochrome c6 complex 20960123, which was found in *Microcoleus vaginatus*. It might be the case that only the cytochrome c6 complex is needed in *Microcoleus vaginatus* because of this redundancy. However, an experiment should still be done to make sure that plastocyanin is truly missing, since there could be errors in the gene annotation.

References

[1] Ravnikar PD, Debus R, Sevrinck J, Saetaert P, McIntosh L: **Nucleotide sequence of a**

**second psbA gene from the unicellular cyanobacterium Synechocystis 6803**.

*Nucleic acids research* 1989, **17**(10):3991.

[2] Lang JD, Haselkorn R: **Isolation, sequence and transcription of the gene encoding the**

**photosystem II chlorophyll-binding protein, CP-47, in the cyanobacterium**

**Anabaena 7120**. *Plant molecular biology* 1989, **13**(4):441-457.

[3] Ferreira KN, Iverson TM, Maghlaoui K, Barber J, Iwata S: **Architecture of the**

**photosynthetic oxygen-evolving center**. *Science* 2004, **303**(5665):1831-1838.

[4] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S *et al*: **Sequence analysis of the genome of the unicellular**

**cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of**

**the entire genome and assignment of potential protein-coding regions**. *DNA Res*

1996, **3**(3):109-136.

[5] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T *et al*: **Complete genomic sequence of the filamentous**

**nitrogen-fixing cyanobacterium Anabaena sp. strain PCC 7120**. *DNA Res* 2001,

**8**(5):205-213; 227-253.

[6] Kaneko T, Nakajima N, Okamoto S, Suzuki I, Tanabe Y, Tamaoki M, Nakamura Y, Kasai F,

Watanabe A, Kawashima K *et al*: **Complete genomic structure of the bloom-forming**

**toxic cyanobacterium Microcystis aeruginosa NIES-843**. *DNA Res* 2007,

**14**(6):247-256.

[7] Sanchez Puerta MV, Bachvaroff TR, Delwiche CF: **The complete plastid genome**

**sequence of the haptophyte Emiliania huxleyi: a comparison to other plastid**

**genomes**. *DNA Res* 2005, **12**(2):151-156.

[8] Oudot-Le Secq MP, Grimwood J, Shapiro H, Armbrust EV, Bowler C, Green BR: **Chloroplast**

**genomes of the diatoms Phaeodactylum tricornutum and Thalassiosira**

**pseudonana: comparison with other plastid genomes of the red lineage**. *Mol*

*Genet Genomics* 2007, **277**(4):427-439.

[9] Nakamura Y, Kaneko T, Sato S, Ikeuchi M, Katoh H, Sasamoto S, Watanabe A, Iriguchi M,

Kawashima K, Kimura T *et al*: **Complete genome structure of the thermophilic**

**cyanobacterium Thermosynechococcus elongatus BP-1**. *DNA Res* 2002,

**9**(4):123-130.

[10] Sugita C, Ogata K, Shikata M, Jikuya H, Takano J, Furumichi M, Kanehisa M, Omata T,

Sugiura M, Sugita M: **Complete nucleotide sequence of the freshwater unicellular**

**cyanobacterium Synechococcus elongatus PCC 6301 chromosome: gene content**

**and organization**. *Photosynth Res* 2007, **93**(1-3):55-67.

[11] Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A,

Ferriera S, Johnson J *et al*: **Patterns and implications of gene gain and loss in the**

**evolution of Prochlorococcus**. *PLoS Genet* 2007, **3**(12):e231.

[12] Welsh EA, Liberton M, Stockel J, Loh T, Elvitigala T, Wang C, Wollam A, Fulton RS, Clifton

SW, Jacobs JM *et al*: **The genome of Cyanothece 51142, a unicellular diazotrophic**

**cyanobacterium important in the marine nitrogen cycle**. *Proc Natl Acad Sci U S A*

2008, **105**(39):15094-15099.

[13] Wales R, Newman BJ, Rose SA, Pappin D, Gray JC: **Characterization of cDNA clones**

**encoding the extrinsic 23 kDa polypeptide of the oxygen-evolving complex of**

**photosystem II in pea**. *Plant molecular biology* 1989, **13**(5):573-582.

[14] Thornton LE, Ohkawa H, Roose JL, Kashino Y, Keren N, Pakrasi HB: **Homologs of plant**

**PsbP and PsbQ proteins are necessary for regulation of photosystem ii activity in**

**the cyanobacterium Synechocystis 6803**. *The Plant cell* 2004, **16**(8):2164-2175.

[15] Jackson SA, Fagerlund RD, Wilbanks SM, Eaton-Rye JJ: **Crystal structure of PsbQ from**

**Synechocystis sp. PCC 6803 at 1.8 A: implications for binding and function in**

**cyanobacterial photosystem II**. *Biochemistry* 2010, **49**(13):2765-2767.

[16] De Las Rivas J, Heredia P, Roman A: **Oxygen-evolving extrinsic proteins**

**(PsbO,P,Q,R): bioinformatic and functional analysis**. *Biochimica et biophysica acta*

2007, **1767**(6):575-582.

[17] Karapetyan NV: **Non-photochemical quenching of fluorescence in cyanobacteria**.

*Biochemistry Biokhimiia* 2007, **72**(10):1127-1135.

[18] Katoh H, Itoh S, Shen JR, Ikeuchi M: **Functional analysis of psbV and a novel c-type**

**cytochrome gene psbV2 of the thermophilic cyanobacterium**

**Thermosynechococcus elongatus strain BP-1**. *Plant & cell physiology* 2001,

**42**(6):599-607.

[19] Swingley WD, Chen M, Cheung PC, Conrad AL, Dejesa LC, Hao J, Honchak BM, Karbach

LE, Kurdoglu A, Lahiri S *et al*: **Niche adaptation and genome expansion in the**

**chlorophyll d-producing cyanobacterium Acaryochloris marina**. *Proc Natl Acad Sci*

*U S A* 2008, **105**(6):2005-2010.

[20] Pagliano C, Chimirri F, Saracco G, Marsano F, Barber J: **One-step isolation and**

**biochemical characterization of a highly active plant PSII monomeric core**.

*Photosynthesis research* 2011, **108**(1):33-46.

[21] Jordan P, Fromme P, Witt HT, Klukas O, Saenger W, Krauss N: **Three-dimensional**

**structure of cyanobacterial photosystem I at 2.5 A resolution**. *Nature* 2001,

**411**(6840):909-917.

[22] Smart LB, McIntosh L: **Expression of photosynthesis genes in the cyanobacterium**

**Synechocystis sp. PCC 6803: psaA-psaB and psbA transcripts accumulate in**

**dark-grown cells**. *Plant Mol Biol* 1991, **17**(5):959-971.

[23] Antonkine ML, Liu G, Bentrop D, Bryant DA, Bertini I, Luchinat C, Golbeck JH, Stehlik D:

**Solution structure of the unbound, oxidized Photosystem I subunit PsaC,**

**containing [4Fe-4S] clusters F(A) and F(B): a conformational change occurs upon**

**binding to photosystem I**. *J Biol Inorg Chem* 2002, **7**(4-5):461-472.

[24] Mann K, Schlenkrich T, Bauer M, Huber R: **The amino-acid sequence of three proteins**

**of photosystem I of the cyanobacterium Fremyella diplosiphon (Calothrix sp PCC**

**7601)**. *Biol Chem Hoppe Seyler* 1991, **372**(7):519-524.

[25] Klukas O, Schubert WD, Jordan P, Krauss N, Fromme P, Witt HT, Saenger W:

**Photosystem I, an improved model of the stromal subunits PsaC, PsaD, and PsaE**.

*J Biol Chem* 1999, **274**(11):7351-7360.

[26] Mayer KL, Shen G, Bryant DA, Lecomte JT, Falzone CJ: **The solution structure of**

**photosystem I accessory protein E from the cyanobacterium Nostoc sp. strain**

**PCC 8009**. *Biochemistry* 1999, **38**(41):13736-13746.

[27] Nyhus KJ, Ikeuchi M, Inoue Y, Whitmarsh J, Pakrasi HB: **Purification and characterization**

**of the photosystem I complex from the filamentous cyanobacterium Anabaena**

**variabilis ATCC 29413**. *J Biol Chem* 1992, **267**(18):12489-12495.

[28] Yu J, Ma PJ, Shi DJ, Li SM, Wang CL: **Homologous comparisons of photosynthetic**

**system I genes among cyanobacteria and chloroplasts**. *Journal of integrative plant*

*biology* 2008, **50**(8):929-940.

[29] Vanselow C, Weber AP, Krause K, Fromme P: **Genetic analysis of the Photosystem I**

**subunits from the red alga, Galdieria sulphuraria**. *Biochimica et biophysica acta*

2009, **1787**(1):46-59.

[30] Kallas T, Spiller S, Malkin R: **Primary structure of cotranscribed genes encoding the**

**Rieske Fe-S and cytochrome f proteins of the cyanobacterium Nostoc PCC 7906**.

*Proc Natl Acad Sci U S A* 1988, **85**(16):5794-5798.

[31] Lucas S, Copeland,A., Lapidus,A., Glavina del Rio,T., Dalin,E., Tice,H., Bruce,D.,

Goodwin,L., Pitluck,S., Sims,D., Meineke,L., Brettin,T., Detter,J.C., Han,C., Larimer,F.,

Land,M., Hauser,L., Kyrpides,N., Ovchinnikova,G., Liberton,M., Stoeckel,J., Banerjee,A.,

Singh,A., Page,L., Sato,H., Zhao,L., Sherman,L., Pakrasi,H.B., Richardson P: **Direct**

**Submission to Swissprot**. 2009.

[32] Yamashita E ZH, Cramer WA.: **Structure of the cytochrome b6f complex: quinone**

**analogue inhibitors as ligands of heme cn.** *JMolBiol* 2007(370):39-52.

[33] Copeland A, Lucas S, Lapidus A, Barry K, Detter JC, Glavina del Rio T, Hammon N, Israni

S, Dalin E, Tice H *et al*: **Direct submission**. In: *Cytochrome b6-f complex subunit petG.*

2006.

[34] Schneider D, Volkmer T, Rogner M: **PetG and PetN, but not PetL, are essential**

**subunits of the cytochrome b6f complex from Synechocystis PCC 6803**. *Research*

*in microbiology* 2007, **158**(1):45-50.

[35] Schneider D, Berry S, Rich P, Seidler A, Rogner M: **A regulatory role of the PetM subunit**

**in a cyanobacterial cytochrome b6f complex**. *J Biol Chem* 2001,

**276**(20):16780-16785.

[36] Navarro JA, Hervas M, De la Rosa MA: **Purification of plastocyanin and cytochrome c6**

**from plants, green algae, and cyanobacteria**. *Methods Mol Biol* 2011, **684**:79-94.

[37] Tanaka M, Haniu M, Yasunobu KT, Rao KK, Hall DO: **Modification of the automated**

**sequence determination as applied to the sequence determination of the**

**Spirulina maxima ferredoxin**. *Biochemistry* 1975, **14**(25):5535-5540.

[38] Cozens AL, Walker JE: **Expression of a gene encoding a novel ferredoxin in the**

**cyanobacterium Synechococcus 6301**. *Biochem J* 1988, **252**(2):563-569.

[39] Hase T, Inoue K, Hagihara N, Matsubara H, Williams MM, Rogers LJ: **Ferredoxin from**

**Aphanothece halophitica, a unicellular blue-green alga: close relationship to**

**ferredoxins from filamentous blue-green algae and phylogenetic implications**. *J*

*Biochem* 1983, **94**(5):1457-1464.

[40] Nakajima M, Sakamoto,T., Wada K: **The complete purification and characterization of**

**three forms of ferredoxin-NADP(+) oxidoreductase from a thermophilic**

**cyanobacterium Synechococcus elongatus**. *Plant Cell Physiol* 2002, **43**((5)):484-493.

[41] Hermoso J, Mayoral, T., Faro, M., Gomez-Moreno, C., Sanz-Aparicio, J., Medina, M.:

**Mechanism of Coenzyme Recognition and Binding Revealed by Crystal Structure**

**Analysis of Ferredoxin-Nadp(+) Reductase Complexed with Nadp(+)**. *JMolBiol*

2002(319):1133.

[42] Bovy A, de Vrieze G, Borrias M, Weisbeek P: **Isolation and sequence analysis of a gene**

**encoding a basic cytochrome c-553 from the cyanobacterium Anabaena SP. PCC**

**7937**. *Plant molecular biology* 1992, **19**(3):491-492.

[43] Groth G, Pohl E: **The structure of the chloroplast F1-ATPase at 3.2 A resolution**. *J Biol*

*Chem* 2001, **276**(2):1345-1352.

[44] Van Walraven HS, Lutter R, Walker JE: **Organization and sequences of genes for the**

**subunits of ATP synthase in the thermophilic cyanobacterium Synechococcus**

**6716**. *Biochem J* 1993, **294 ( Pt 1)**:239-251.

[45] Curtis SE: **Genes encoding the beta and epsilon subunits of the proton-translocating**

**ATPase from Anabaena sp. strain PCC 7120**. *J Bacteriol* 1987, **169**(1):80-86.

[46] Stocker A, Keis S, Vonck J, Cook GM, Dimroth P: **The structural basis for unidirectional**

**rotation of thermoalkaliphilic F1-ATPase**. *Structure* 2007, **15**(8):904-914.

[47] McCarn DF, Whitaker RA, Alam J, Vrba JM, Curtis SE: **Genes encoding the alpha,**

**gamma, delta, and four F0 subunits of ATP synthase constitute an operon in the**

**cyanobacterium Anabaena sp. strain PCC 7120**. *J Bacteriol* 1988, **170**(8):3448-3458.

[48] Yagi H, Konno H, Murakami-Fuse T, Isu A, Oroguchi T, Akutsu H, Ikeguchi M, Hisabori T:

**Structural and functional analysis of the intrinsic inhibitor subunit epsilon of**

**F1-ATPase from photosynthetic organisms**. *Biochem J* 2010, **425**(1):85-94.

[49] Sazuka T, Ohara O: **Towards a proteome project of cyanobacterium Synechocystis**

**sp. strain PCC6803: linking 130 protein spots with their respective genes**.

*Electrophoresis* 1997, **18**(8):1252-1258.

[50] Pogoryelov D, Yildiz O, Faraldo-Gómez JD, Meier T: **High-resolution structure of the**

**rotor ring of a proton-dependent ATP synthase**. *Nat Struct Mol Biol* 2009,

**16**(10):1068-1073.

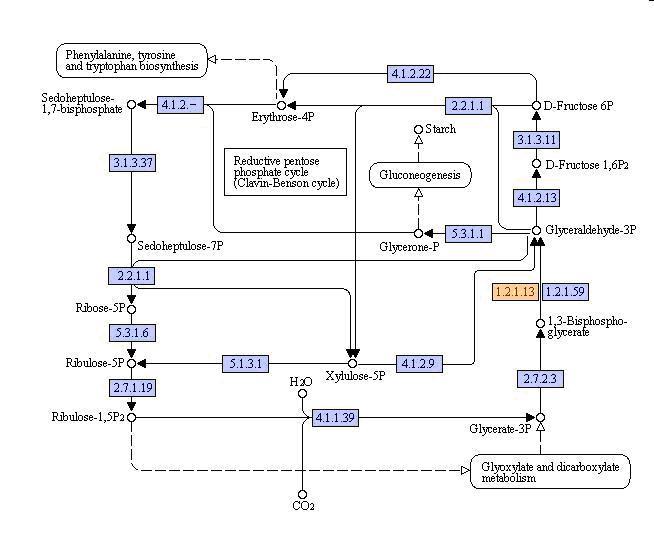
[51] Cozens AL, Walker JE, Phillips AL, Huttly AK, Gray JC: **A sixth subunit of ATP synthase,**

**an F(0) component, is encoded in the pea chloroplast genome**. *The EMBO journal*

1986, **5**(2):217-222.

# Calvin Cycle

The Calvin Cycle was annotated for *Microcoleus vaginatus* PCC 9802. The MetaCyc Calvin Cycle pathway was consulted to see which enzymes were specific to the pathway, and genes were found in *Microcoleus vaginatus* that coded for those enzymes. The pathway is important since it is part of photosynthesis and is essential for the organism to survive. The Calvin Cycle is the major carbon fixation pathway for the organism. The cycle goes through 6 turns, taking in 6 CO2 molecules, in order to produce one hexose sugar. In this cycle, one CO2 molecule at a time is added to the acceptor molecule D-ribulose-1,5-bisphosphate (RuBP) generating two molecules of 3-phospho-D-glycerate. The 3-phospho-D-glycerate is then passed through a cyclic series of reactions in which a hexose sugar is produced and the acceptor molecule (RuBP) is regenerated. Enzyme 1.2.1.13 was missing in the automated annotation, but all functions were found in the final annotation.

**Automated Annotation**

S Fig 1: Calvin Cycle Automated Annotation. Blue boxes correspond to M. vaginatus genes with the predicted EC#, Orange boxes correspond to functions that exist in other genomes but were not predicted in M. vaginatus (by KEGG). All the genes related to the Calvin Cycle were checked and annotated below.

**3-phosphoglycerate kinase** (EC: 2.7.2.3, Gene OID: 2505166234)

A PDB BLAST search of the gene gave as a first result an article on the structure of 3-phosphoglycerate kinase in *Bacillus stearothermophilus* [3] (PDB: 1PHP). A reverse BLAST of this hit gave gene 2505166234 as the top hit. All active sites matched except for sites 233, 292 and 323. In the paper, the sites were Tyr 223, Ala 292 and Asp 323. Microcoleus vaginatus had at these sites Phe 223, Gly 292 and Glu 323. The change in site 223 seems to be justifiable because Tyr and Phe are similar in structure.

A multiple sequence alignment [21] was done with the *Microcoleus vaginatus* FASTA sequence for the gene and other FASTA sequences from IMG/ER of cyanobacteria genes that code for the enzyme . The alignment showed that Phe is almost completely conserved for site 223 compared to other cyanobacteria, with only one other organism having Tyr. The change in site 292 seems justifiable as well. The alignment showed that Gly is completely conserved in other cyanobacteria for site 292. Finally, the alignment showed complete conservation of Glu for site 323.

**Fructose-1,6-bisphosphatase**  EC: 3.1.3.11

OID: 2505169097

**Fructose-1,6-bisphosphatase/Sedoheptulose Bisphosphatase** EC: 3.1.3.37 and 3.1.3.11

OID: 2505170007

There were two hits in IMG ER for this enzyme number. A BLAST search using gene 2505169097 was done on UniProt [22], and the results were narrowed down to reviewed results for cyanobacteria. These results were then split into hits for EC 3.1.3.11 only, hits for EC 3.1.3.11 and 3.1.3.37, and hits for EC 3.1.3.37 only. A multiple sequence alignment [21] of each of these subsets was done against the FASTA sequences of genes 2505169097 and 2505170007. The results suggested that gene 2505169097 was most closely correlated to EC 3.1.3.11, while gene 2505170007 was most closely correlated to being bifunctional for EC 3.1.3.11 and 3.1.3.37. As a final check, a phylogenetic tree [18] was made with the FASTA sequences used for the alignments mentioned previously. The resulting tree showed that gene 2505170007 clustered with the bifunctional genes, while gene 2505169097 clustered with the genes for EC 3.1.3.11.

**Triosephosphate Isomerase** EC: 5.3.1.1

Gene OID: 2505168240

A PDB BLAST search of the gene gave a paper on the structure of triosephosphate isomerase in *Bacillus stearothermophilus* [4] (PDB: 1BTM). All active site residues mentioned in the paper were identical to the ones in *Microcoleus vaginatus* except for site 12.

According to the paper, residue 12 is usually an Asn, but in *Bacillus stearothermophilus*, it has been mutated to His 12. In *Microcoleus vaginatus*, it is Tyr 12. This mismatch could be explained by the fact that His and Tyr both have aromatic rings. Additionally, a multiple sequence alignment [21] was done against all cyanobacteria genes in IMG/ER for EC 5.3.1.1 and against reviewed cyanobacteria gene hits from a UniProt BLAST [22]. This revealed that His and Asn commonly appear at site 12, along with Tyr and Phe. It seemed that cyanobacteria were partial to the aromatic amino acids Tyr, His and Phe. From the alignment, it seems that Tyr 12 is a valid active site residue. The paper found from the PDB BLAST was published in 1995, so it might just be that these other amino acids are valid for the active site, but no sequences using these positions were known back at the time the paper was published.

**Ribose 5-Phosphate Isomerase** EC: 5.3.1.6

Gene OID: 2505165655

BLAST searches within IMG/ER and UniProt [22] both gave ribose 5-phosphate isomerases as the top hits. A PDB BLAST search gave as a first result an article on the structure of ribose-5-phosphate isomerase in *Pyrococcus horikoshii* [6] (PDB: [1LK5]).

The active sites mentioned in the article were checked against those of *Microcoleus vaginatus*, and they all matched except for sites 100, 106 and 168.

The sites were R100, M106 and D168 in *Pyrococcus horikoshii*. In *Microcoleus vaginatus*, the sites were G100, R106 and D168. In the article, there was no mention of the functions of sites 106 or 168, and it said that site 100 was not completely conserved. Also, a multiple sequence alignment was shown in the article. This alignment showed that the aforementioned sites were not completely conserved, and the amino acids found in *Microcoleus vaginatus* were found in some of the other organisms used for the alignment. To double-check, a multiple sequence alignment [21] was also done against other similar genes from cyanobacteria found from the UniProt [22] database with a BLAST search. Site 100 was almost completely conserved along cyanobacteria as being a G, except for one cyanobacterium that was a T. Site 106 had many R's among other cyanobacteria, as well as Q and a single A. Site 168 had mostly A's among cyanobacteria, a couple D's, and one T. The few outliers could be explained as not actually being the correct enzyme, because some cyanobacteria listed multiple genes for EC 5.3.1.6, which could be a mistake due to automatic annotation.

**Glyceraldehyde-3-phosphate Dehydrogenase** EC 1.2.1.12 / 1.2.1.13 / 1.2.1.59

Gene OID: 2505167745 and 2505170636

EC 1.2.1.13 performs the same conversion of DPG to G3P as EC 1.2.1.12 and EC 1.2.1.59 do, except it consumes NADPH in the process, while 1.2.1.12 consumes NADH and 1.2.1.59 can use both. 1.2.1.13 performs this step in the Calvin Cycle, yet no genes in any cyanobacteria in IMG ER were annotated to have this activity. *Synechocystis* PCC 6803's 1.2.1.12 and 1.2.1.13 sequences were found in an article [13], and then BLASTed against the genome of M. vaginatus. The top homolog been predicted to be an enzyme with EC 1.2.1.59. A subsequent multiple sequence alignment [21] and phylogenetic tree analysis [18] of all genes in cyanobacteria containing pfam02800 and pfam00044 showed two distinct subgroups and several outliers. The gene in M. vaginatus (2505167745) that had been predicted to be 1.2.1.59 clustered with the 12's, while the only other homolog in the genome was one of the four outliers (2505170636). There is a structure of 1.2.1.13 from *Synechococcus elongatus* PCC 7942 [11] (PDB: 2D2I). The active sites do seem to match up very well with the predicted active sites in 1.2.1.13. Almost all 1.2.1.12 structures on PDB come from eukaryotes, archaea, and very distantly related bacteria, and no cyanobacterial data on this enzyme exist.When we built a phylogenetic tree [18] using the two M. vaginatus sequences and confirmed 1.2.1.13/59/12 sequences from UniProt [22], the result was that 2505170636 clustered with a group of 1.2.1.12's and 2505167745 was in a cluster that contained all three enzymes. The ambiguity of the results suggests that this should be verified in the lab. A number of assays exist to measure the concentration of NADPH/NADP+; most of them use an enzyme cycling reaction to improve spectrophotometric detection; a similar assay exists for NADH/NAD+.

**Ribulose-phosphate 3-epimerase** EC: 5.1.3.1

Gene OID: 2505170722

This gene most likely encodes Ribulose-phosphate 3-epimerase EC 5.1.3.1 due to the strong homology it bears to verified ribulose-phosphate 3-epimerases found in PDB [26] (PDB: 1TQJ) and Swissprot databases [12] (NCBI: P74061) with EC 5.1.3.1. A reverse BLAST search gave gene 2505170722 as the top hit.

**RuBisCo Activase**

Gene OID: 2505170762

This gene most likely encodes RuBisCo Activase. In addition to the strong homology it bears to verified RuBisCo Activase found in Swissprot databases [14] (NCBI: Q06721), its dominant pfam AAA is associated with activase and is thought to be used for substrate recognition. This gene is not included in the pathway because it is not an enzyme that functions in the pathway.

**Ribulose bisphosphate carboxylase/oxygenase large chain** EC: 4.1.1.39

Gene OID: 2505165766

This gene most likely encodes the large chain of RuBisCo (EC 4.1.1.39) due to the strong homology it bears to verified chains of RuBisCo found in PDB [10] (PDB: 1UPM) and Swissprot databases [2] (NCBI: Q10WH6) with EC 4.1.1.39.

**Ribulose bisphosphate carboxylase/oxygenase small chain** EC: 4.1.1.39

Gene OID: 2505165768

This gene most likely encodes the small chain of RuBisCo due to the strong homology it bears to verified small chains of RuBisCo found in PDB [17] (PDB: 1RSC) and Swissprot databases [1] (NCBI: P54206) with EC 4.1.1.39.

**Transketolase** EC: 2.2.1.1

Gene OID: 2505170664

This gene most likely encodes for the 2.2.1.1 reaction enzyme due to homology and 100% active site matching with transkelotases found in PDB [25] (PDB: 1AYO) and Swissprot databases [9] (NCBI: P54206) with EC 2.2.1.1.

**2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase** EC: 4.1.2.-

Gene OID: 2505167217

This gene most likely encodes for the 4.1.2.- reaction enzyme due to homology and close active site matching with APO class II Aldolase HPCH in PDB [19] (PDB: 2V5J) with EC 4.1.2.-. 7/9 active sites matched with the exceptions Q147 to H147 and A42 to Y42. Phylogenic analysis suggests that similarity to another aldolase (retrieved from PDB blast) with EC 4.1.2.20 is not significant. Although 4.1.2.- enzymes can undergo 4.1.2.13 reactions there is weak phylogenetic grouping associated with this M. vaginatus gene and other 4.1.2.13 enzymes. Lab work might be needed to confirm this prediction.

**Uridine Kinase** EC: 2.7.1.19

Gene OID: 2505171016 and 2505171061

Genes 2505171016 and 2505171061 were both predicted to code for enzyme 2.7.1.19 (uridine kinase). Both genes were predicted, with a high degree of numerical reliability, to have the same COG ID and Pfam number. NCBI BLAST searches on the Swissprot Database [20] (NCBI: P37101) for both genes gave, also with a high degree of numerical reliability, other phosphoribulokinase enzymes with the same EC numbers as top hits. PDB BLAST Searches for genes 2505171016 and 2505171061 [23] (PDB: 3ASY) did not give any protein structures with matching EC numbers. The first PDB BLAST hit for gene 2505171016 was [3ASY], a ligand-free structure of uridine kinase from the gram-negative bacteria *Thermus thermophilus* (HB8 strain). The article for this structure was not accessible, and the rest of the PDB hits were of human uridine kinase enzyme structures. All the top hits for gene 2505171061 were of human uridine kinase enzyme structures, and the first non-human protein structure for gene 2505171061 was the same as the top hit for gene 2505171016, [3ASY]. Since the data above did not provide a solid conclusion, the gene ortholog neighborhoods for both genes were analyzed. Analysis of the neighborhoods showed gene 2505171061 to have more conserved genes and clusters that seemed to be more relevant to the Calvin Cycle (like Ferredoxin--NADP(+) reductase) than gene 2505171016. Based on the data above, we predict that gene 2505171061 is more likely to code for enzyme 2.7.1.19 than gene 2505171016.

**Fructose-Biphosphate Aldolase** EC: 4.1.2.13

Gene OID: 2505166282 and 2505169777

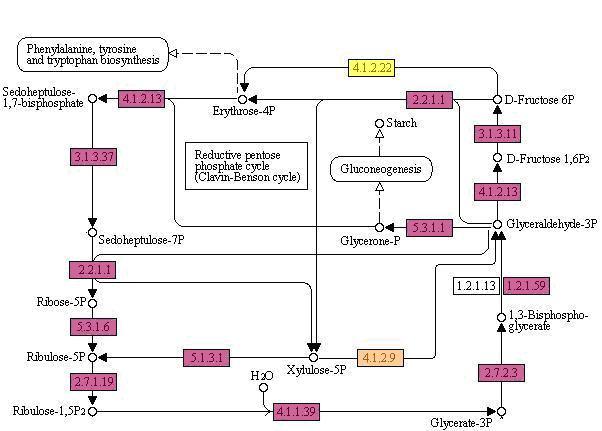
Genes 2505166282 (fructose-bisphosphate aldolase, class II, Calvin cycle subtype) and 2505169777(DhnA-type fructose-1,6-bisphosphate aldolase and related enzymes) were both predicted to code for enzyme 4.1.2.13 (fructose-bisphosphate aldolase, abbreviated as FBPA). FBPA catalyzes the reversible cleavage of D-fructose-1,6-bisphosphate (FBP) to dihydroxyacetone phosphate (DHAP) and D-glyceraldehyde-3-phosphate (G3P), a key step of the classical Embden-Meyerhof-Parnas glycolytic pathway.

NCBI Blast Searches against the Swissprot Database, with a high degree of numerical reliability, matched gene 2505166282 with class II fructose-1,6-bisphosphate aldolase enzymes [7] (NCBI: Q9XDP3), and matched gene 2505169777 with fructose-bisphosphate enzymes [24] (NCBI: P0A992). IMG/ER Gene Detail results, also with a high degree of numerical reliability, matched gene 2505166282 with a COG ID of COG0191 and the F\_bP\_aldolase Pfam domain, but matched gene 2505169777 with a COG ID of COG1830 and the DeoC Pfam domain. These results were unusual since both genes were matched with EC number 4.1.2.13.

PDB BLAST Searches for each of the genes provided more clarity into the matter. Gene 2505166282, with high numerical reliability and 100% active site matches, was matched with protein structure 3GAK [5]. An article on this structure specified that this enzyme, located in disease-causing protozoan *Giardia lamblia*, belongs to the class II aldolase family which employs a Zinc (2+) cofactor, rather than a lysine active site like the class I aldolases do.

Gene 2505169777, with high numerical reliability and 100% active site matches, was matched with protein structure 1OJX [15]. An article on this structure specified that this enzyme, located in the hyperthermophile *Trichomonas tenax*, belongs to the recently-identified third type of aldolase class, class IA for archaeal FBPA (fructose-bisphosphate aldolase). While the archaeal FBPA (IA) catalyze the same reactions and have the same fold as FBPA I, FBPA IA and FBPA do not share the same overall sequence identity.

The KEGG database identifies enzyme 4.1.2.13 with having a zinc cofactor. This probably implies that enzyme 4.1.2.13 is a class II FBPA, similar to the one located with *Giardia lamblia*. However, both genes 2505166282 and 2505169777 accurately for some type of FBPA. Both types of FBPA are predicted to share the same function, but employ different structures, but thorough research has not been conducted to prove this as of yet. To further research this matter, we could conduct research comparing the functional and structural differences between FBPA II and FBPA IA.

**Comparison with other Genomes**

Genes found in some of the selected genomes

S Fig. 2: Comparison of genes found in M. vaginatus versus genes found in other cyanobacteria by automatic annotation of the Calvin Cycle

Genes from Microcoleus vaginatus were found in most of the other cyanobacteria genomes. There were no new genes found in other cyanobacteria that were not in Microcoleus Vaginatus.

**Functional Studies**

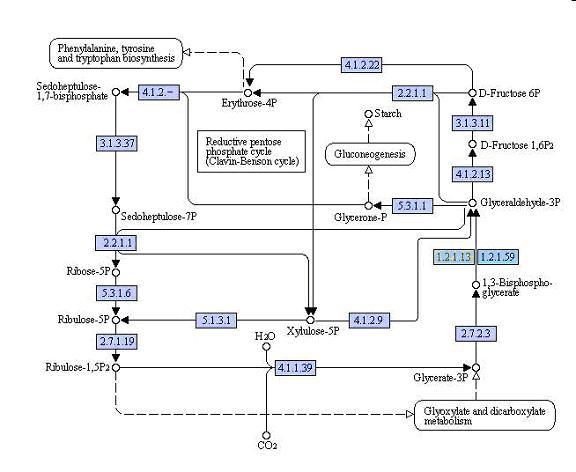
Functional studies could be done for the genes associated with EC 3.1.3.37 and 3.1.3.11. Gene 2505170007 should be tested to see if it is actually bifunctional. Gene 2505169097 should be repressed and it should be seen if EC 3.1.3.37 is still expressed. Gene 2505170007 should then be repressed separately as well to see if EC 3.1.3.11 is not expressed. If both of these conditions are met, this means gene 2505170007 is bifunctional.

Additionaly, EC 3.1.3.11 takes fructose 1,6-bisphosphate as a substrate, whereas EC 3.1.3.37 takes sedoheptulose 1,7-bisphosphate as its substrate. Therefore, experiments could be done where only one gene is repressed and involving only one of the substrates at a time, to confirm if the genes code for their specified enzymes.

Functional studies should also be done for the genes associated with EC 1.2.1.12, 1.2.1.13 and 1.2.1.59. The functions of genes 2505167745 and 2505170636 were ambiguous. Their function can be determined by running a test where equal amounts of NADPH and NADH are provided to the system. An assay [4] should be used to determine the concentrations of NADPH/NADP+ and NADH/NAD+ after the organism is allowed to go through the Calvin cycle. If there are high concentrations of NADP+, the expressed gene codes for EC 1.2.1.13. If there are high concentrations of NAD+, the expressed gene codes for EC 1.2.1.12. If there are similar levels of NADP+ and NAD+, then the expressed gene codes for EC 1.2.1.59.

Functional studies can be done for genes 250516282 and 2505169777, which correspond to EC 4.1.2.13. The genes are both predicted to code for FBPA, but there exists different classes of FBPAs. Studies with other organisms like e. coli could be done to try and first S Figure out the differences between the different classes of FBPAs, and then observations could be made about the FBPAs in M. vaginatus.

Finally, functional studies should be done with gene 2505167217, which was predicted to code for EC 4.1.2.-. It should be checked to see if this gene can indeed code for EC 4.1.2.20 or EC 4.1.2.13 in experiment by seeing if those enzymes are expressed in the absence of all other genes that also code for the aforementioned enzymes. Such an experiment could show that gene 2505167217 is either bifunctional or improperly annotated.

**Final Annotation** 

S Fig. 3: Genes found in Calvin Cycle (blue) and maybe found in Calvin Cycle (light blue) for M. vaginatus

All of the genes in blue for the Calvin Cycle were found to be present, and they were confirmed to code for the specificed enzymes. The missing enzyme 1.2.1.13 was found to possible exist, while the enzyme 1.2.1.59 might or might not exist.

**Discussion**

*Microcoleus vaginatus* was found to have a complete Calvin cycle. The traditional Calvin cycle pathway was present, along with an alternative pathway through erythrose-4P, sedoheptulose-1,7-bisphosphate, sedoheptulose-7P, and ribose-5P to ribulose-5P.

*Microcoleus vaginatus* was found to have a bi-functional gene (2505170007) that coded for both EC 3.1.3.37 and enzyme 3.1.3.11. EC 3.1.3.11 is used in the traditional Calvin Cycle pathway, while EC 3.1.3.37 is used in the alternative pathway. It is uncertain why another gene (2505169097) that only codes for EC 3.1.3.11, since this gene is not needed if gene 2505170007 can already code for EC 3.1.3.11. An experiment could be done where gene 2505169097 is repressed and observations are taken to see if there are any adverse effects on the organism.

There were two genes, 250516282 and 2505169777, that both coded for EC 4.1.2.13, although they could code for different classes of the same enzyme. This result is inconsequential in that the Calvin cycle still works. The different class enzymes have differing structures, but they have the same function, and in the big scheme of the pathway, the enzymes work as long as they have the correct function.

Additionally, microcoleus vaginatus was found to have two homologous genes that could code for either EC 1.2.1.12, 1.2.1.13 or 1.2.1.59. These enzymes were substrate-specific to NADH, NADPH and both NADH or NADPH respectively. It would be predicted that the two genes code for EC 1.2.1.13 and 1.2.1.59, because both the enzymes accept NADPH as a substrate, and it is known that NADPH is the primary electron carrier for the Calvin cycle. If it turns out the one of the genes actually does encode EC 1.2.1.12 however, this would be interesting because NADH is not used for the Calvin cycle. Such a finding would merit further study on the Calvin cycle pathway in *Microcoleus vaginatus*.

References

[1] Amichay, D., Levitz, R. and Gurevitz, M.: Construction of a Synechocystis PCC6803

mutant suitable for the study of variant hexadecameric ribulose bisphosphate

carboxylase/oxygenase enzymes. Plant Mol. Biol. 23 (3), 465-476 (1993)

[2] Copeland, A., Lucas, S., Lapidus, A., Barry, K., Detter, J.C., et al.:

Swissprot Direct Submission (Jun 2006)

[3] Davies, G.J., Gamblin, S.J., Littlechild, J.A., et al.: Structure of the ADP complex of

the 3-phosphoglycerate kinase from *Bacillus stearothermophilus* at 1.65 A. (1994) Acta Crystallogr., Sect. D 50: 202-209.

[4] Delboni, L.F., Mande, S.C., Rentier-Delrue, F., et al.: Crystal structure of

recombinant triosephosphate isomerase from Bacillus stearothermophilus. An analysis

of potential thermostability factors in six isomerases with known three-dimensional

structures points to the importance of hydrophobic interactions. (1995) Protein Sci. 4:

2594-2604.

[5] Galkin A, Li Z, Li L, Kulakova L, Pal LR, Dunaway-Mariano D[,](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Galkin%20A%22%5BAuthor%5D) Herzberg O.:

Structural Insights into the Substrate Binding and Stereoselectivity of Giardia

Fructose-1,6- bisphosphate Aldolase. (2009)Biochemistry 48: 3186-3196

[6] Ishikawa, K., Matsui, I., Payan, F, et al.: A hyperthermostable D-ribose-5-

phosphate isomerase from Pyrococcus horikoshii characterization and

three-dimensional structure. (2002) Structure 10: 877-886.

[7] Joardar,V., Ehling-Schulz,M., Smith,S.C., Scherer,S. and Potts,M.; Direct Submission

to Swissprot. Submitted (??-MAY-1997)

[8] Kaneko,T., Nakajima,N., Okamoto,S., Suzuki,I., Tanabe,Y., et al.Complete genomic

structure of the bloom-forming toxic cyanobacterium Microcystis aeruginosa

NIES-843. DNA Res. 14 (6), 247-256 (2007)

[9] Kaneko,T., Nakamura,Y., Wolk,C.P., Kuritz,T., Sasamoto,S., et al.: Complete

genomic sequence of the filamentous nitrogen-fixing cyanobacterium Anabaena sp.

strain PCC 7120. DNA Res. 8 (5), 205-213 (2001)

[10] Karehabadi, S., Taylor, T.C., Andersson, I.: Calcium supports loop closure but not

catalysis in Rubisco. J. Mol.Biol. 334: 65 (2003)

[11] Kitatani, T., Nakamura, Y., Wada, K., Kinoshita, T., Tamoi, M., Shigeoka, S., Tada,

T.:Structure of NADP-dependent glyceraldehyde-3-phospate dehydrogenase from Synechoccus PCC 7942 complexed with NADP. (2006) Acta Crystallogr.,Sect.F 62: 315-319

[12] Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., et al.:

The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature

390 (6657), 249-256 (1997)

[13] Kurian D, Jansen T, Maenpaa P.: Proteomic analysis of heterotrophy in

Synechochystis sp. PCC 6803. Proteomics, 2006 Mar: 6(5): 1483-9

[14] Li, L.A., Gibson,J.L. and Tabita,F.R.: The Rubisco activase (rca) gene is located

downstream from rbcS in Anabaena sp. strain CA and is detected in other

Anabaena/Nostoc strains. Plant Mol. Biol. 21 (5), 753-764 (1993)

[15] Lorentzen E, Pohl E, Zwart P, Stark A, Russell RB, Knura T, Hensel R, Siebers B.

Crystal Structure of an Archaeal Class I Aldolase and the Evolution of (βα)8

Barrel Proteins. (2003) J.Biol.Chem. 278: 47253

[16] NADPH and NADH Assay Kits <http://www.abcam.com/NADP-NADPH-Assay-

Kit-ab65349.html> and <http://www.abcam.com/nad-nadh-assay-kit-

ab65348.html>

[17] Newman, J., Gutteridge, S.: Structure of an effector-induced inactivated state of

ribulose 1,5-bisphosphatecarboxylase/oxygenase: the binary complex between enzyme

and xylulose 1,5-bisphosphate. Structure 2: 495-502 (1994)

[18] Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res.

2008 Jul 1;36(Web Server issue):W465-9. Epub 2008 Apr 19. (PubMed)

[19] Rea, D., Fulo, V., Bugg, T.D.H., Roper, D.I.:Structure and mechanism of HpcH: a

metal ion dependent class II aldolase from the homoprotocatechuate degradation

pathway of Escherichia coli. J.Mol.Biol. 373: 866 (2007)

[20] Su,X. and Bogorad,L.: A residue substitution in phosphoribulokinase of

Synechocystis PCC 6803 renders the mutant light-sensitive. J. Biol. Chem. 266

(35), 23698-23705 (1991)

[21] T-Coffee: A novel method for fast and accurate multiple sequence alignment. (2000) J Mol  
 Biol. 302(1):205-17.

[22] The UniProt Consortium. Ongoing and future developments at the Universal Protein  
 Resource Nucleic Acids. Res. 39: D214-D219 (2011).

[23] Tomoike F, Nakagawa N, Kuramitsu S, Masui R.: A Single Amino Acid Limits the Substrate  
 Specificity of Thermus thermophilus Uridine-Cytidine Kinase to Cytidine. (2011)  
 Biochemistry 50: 4597-4607

[24] Welch,R.A., Burland,V., Plunkett,G. III, Redford,P., Roesch,P.,et al.: Extensive mosaic  
 structure revealed by the complete genome sequence of uropathogenic   
 Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. 99 (26), 17020-17024 (2002)  
  
[25] Wikner, C., Nilsson, U., Meshalkina, K., Udekwu, C., Lindgvistm Y., Schneider, G.:  
 Identification of catalytically important residues in yeast transkelotase. (1997)  
 Biochemistry 36: 15643-15649

[26] Wise, E.L., Akana, J., Gerlt, J.A., Rayment, I.: Structure of D-ribulose 5-phosphate

3-epimerase from Synechocystis to 1.6 A resolution. (2004) Acta Crystallogr.,Sect.D 60:

1687-1690

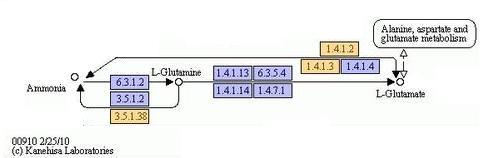
# Nitrogen Metabolism

## Ammonium Assimilation

Although atmospheric nitrogen (N2) is most likely not fixed by *M. vaginatus*, the cyanobacteria still metabolises nitrogen in the form of ammonia for its necessary functions. Ammonium may be one of those key nitrogen sources. In this annotation IMG/ER automated predictions as well as manual predictions of key enzymes, transport proteins, and regulators are studied.

Ammonium, the most reduced form of nitrogen, is a crucial step in the nitrogen assimilation process. “In cyanobacteria, after transport by speciﬁc permeases, ammonium is incorporated into carbon skeletons by the sequential action of glutamine synthetase (GS) and glutamate synthase (GOGAT). Two types of GS (GSI and GSIII) and two types of GOGAT (ferredoxin-GOGAT and NADH–GOGAT) have been characterized in cyanobacteria. The carbon skeleton substrate of the GS–GOGAT pathway is 2-oxoglutarate that is synthesized by the isocitrate dehydrogenase (IDH). In order to maintain the C–N balance and the amino acid pools homeostasis, ammonium assimilation is tightly regulated. The key regulatory point is the GS, which is controlled at transcriptional and posttranscriptional levels. The transcription factor NtcA plays a critical role regulating the expression of the GS and the IDH encoding genes. In the unicellular cyanobacterium Synechocystis sp. PCC 6803, NtcA controls also the expression of two small proteins (IF7 and IF17) that inhibit the activity of GS by direct protein–protein interaction. Cyanobacteria perceive nitrogen status by sensing the intracellular concentration of 2-oxoglutarate, a signaling metabolite that is able to modulate allosterically the function of NtcA, in vitro. In vivo, a functional dependence between NtcA and the signal transduction protein PII in controlling NtcA dependent genes has been also shown.” [[11]](http://198.128.29.183/genome/converter.php#ref-11)

**Automated Annotation**



S Fig 1. Kegg Illustration of Ammonium Assimilation & glutamate synthesis pathway of Nitrogen Metabolism. IMG/ER automated predictions are indicated by color: blue indicates a gene prediction in *M. vaginatus*, orange indicates a gene prediction in other genomes but not in *Microcoleus,* and white indicates no prediction was made. The original pathway map can be found at<http://www.genome.jp/kegg/pathway/map/map00910.html>.

**Ammonium Transport Proteins (amt1 & amt2)**

Gene OID: 2505169989 TC: 1.A.11.2.3

This is a probable Ammonium permease encoding gene inferred from homology with known Synechocystis sp. PCC 6803 amt encoding genes found in NCBI (P54147, 58% identity, E-value 0) [[10]](http://198.128.29.183/genome/converter.php#ref-10). In addition, this gene was found in the same gene neighborhood as the predicted amt2.

Gene OID: 2505169985 TC: 1.A.11.2.3

This is a probable Ammonium permease encoding gene inferred from homology with verified amt encoding genes found in NCBI (Q07429, 49% identity, 2.00E-123) [[18]](http://198.128.29.183/genome/converter.php#ref-18). This gene was also found in the same neighborhood as the predicted amt1.

**Glutaminase**

Gene OID: 2505167437 EC 3.5.1.2

This gene is predicted to code for glutaminase due to strong homology with experimentally verified glutaminase enzymes and 100% active site conservation. The top verified homolog with EC 3.5.1.2 and active site information was found in PDB (O31465, 32% identity, E-value: 6.41E-41) [[4]](http://198.128.29.183/genome/converter.php#ref-4).

**Glutamine synthetase type I**

Gene OID: 2505167133 EC: 6.3.1.2

This gene is predicted to code for L-glutamine synthetase type I due to strong homology with experimentally verified synthetase type I enzymes and 100% active site conservation. Strong homologs and active sites coding for EC 6.3.1.2 were found in PDB (P0A590, 57% identity, E-value 9.94E-165) [[9]](http://198.128.29.183/genome/converter.php#ref-9) and Swissprot (P77961, 81% identity, E-value 0) [[15]](http://198.128.29.183/genome/converter.php#ref-15).

**Glutamine synthetase type III**

Gene OID: 2505169287 EC: 6.3.1.2

This gene is predicted to code for type-3 glutamate--ammonia ligase type III due to strong homology with experimentally verified synthetase type III enzymes and 100% active site conservation. Strong homologs and active sites coding for EC 6.3.1.2 were found in PDB (Q5LGP1, 43% identity, E-value 5.91515E-165) [[16]](http://198.128.29.183/genome/converter.php#ref-16) and Swissprot (Q54WR9, 46% identity, E-value 0) [[6]](http://198.128.29.183/genome/converter.php#ref-6).

**Glutamate synthase (NADH/NADPH) small subunit**

Gene OID: 2505168845 EC: 1.4.1.13/1.4.1.14

This gene is predicted to code for the small subunit of glutamate synthase (NADH/NADPH) due to strong homology with experimentally verified synthase enzymes. The top verified homolog with EC 1.4.1.13/1.4.1.14 was found in Swissprot (O34399, 60% identity, E-value 0) [[1]](http://198.128.29.183/genome/converter.php#ref-1).

**Glutamate synthases (NADH/NADPH) large subunit**

Gene OID: 2505169423 EC: 1.4.1.13/1.4.1.14

This gene is predicted to code for the large subunit of glutamate synthase (NADH/NADPH) due to strong homology and close active site matching with experimentally verified synthase enzymes. The top verified homolog with active site information was found in PDB (Q05755, 44% identity, E-value 0) [[2]](http://198.128.29.183/genome/converter.php#ref-2). Active site mismatches included Q1107-V, G1093-A, and G857-A out of 16 active sites. Further research may be necessary if a thorough investigation of these substitutions is crucial.

**Asparagine synthase (glutamine-hydrolysing)**

Gene OID: 2505166041 & 2505168423 EC: 6.3.5.4

These are possible Asparagine synthase (glutamine-hydrolysing) encoding genes inferred from homology. A weak homolog from *Bacillus subtilis* (P54420) [[19]](http://198.128.29.183/genome/converter.php#ref-19) was found in Swissprot (2505166041: 24% identity, E-value 7.00E-42; 2505168423: 22% identity, E-value 2.00E-27). In addition, a multiple sequence alignment with T-coffee using different asapargine synthases predicted in cyanobacteria showed that only certain domains were conserved (in terms of amino acid properties).

**Glutamate synthase (ferredoxin dependent)**

Gene OID: 2505170210 EC: 1.4.7.1

This gene is predicted to code for ferredoxin dependent glutamate synthase due to strong homology and close active site matching with experimentally verified enzymes with EC 1.4.7.1. The top verified homolog with active site information was found in PDB (P55038, 65% identity, E-value 0) [[22]](http://198.128.29.183/genome/converter.php#ref-22). The only active site mismatch (out of 16 sites) was Ser1129-T, which most likely does not drastically affect the protein’s function.

**Glutamate dehydrogenase**

Gene OID: 2505169473 EC: 1.4.1.3

This is a probable GDH encoding gene inferred from homology and active site conservation. Strong homologs with EC 1.4.1.2, 1.4.1.3 and 1.4.1.4 were found in NCBI and PDB databases, which suggests that the protein can most likely handle both NAD+ and NADP+ as cofactors. Active sites matched 100% in two different homologs in PDB with EC 1.4.1.2 (O74024, 46% identity, E-value 1.28E-102) [[13]](http://198.128.29.183/genome/converter.php#ref-13)and 1.4.1.3 (P96110, E=3.54703E-91, 43% identity) [[8]](http://198.128.29.183/genome/converter.php#ref-8).

**global nitrogen regulator NtcA**

Gene OID: 2505167256

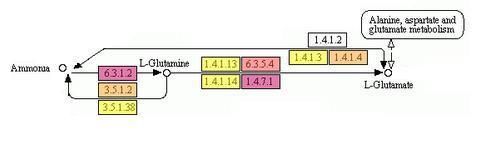
This is a probable global nitrogen regulator NtcA inferred from homology with experimentally verified homologs found in NCBI (P0A4U6, 87% identity, E-value 1.00E-141) [[17]](http://198.128.29.183/genome/converter.php#ref-17). In addition to this homology, 2-oxoglutarate binding site conservation was observed with another homology in PDB (P0A4U6) [[20]](http://198.128.29.183/genome/converter.php#ref-20).

**Nitrogen regulatory protein PII**

Gene OID: 2505170043

This is a probable PII encoding gene inferred from homology and binding site (with 2-oxoglutarate) conservation. Strong homologs were found in PDB (O30794, 92% identity, E-value 2.00E-69) [[7]](http://198.128.29.183/genome/converter.php#ref-7) as well as known binding sites (P0A3F4) [[5]](http://198.128.29.183/genome/converter.php#ref-5), which were conserved in this predicted protein.

**Comparison with other genomes**





S Fig 2. IMG/ER illustration of enzyme conservation among 68 different cyanobacterial genomes in Kegg pathway.

As a result of comparisons with other genomes, enzymes encoding Glutamin-(asparagin-)ase EC 3.5.1.38 most likely do not exist in *M. vaginatus* due to poor homology (via BLAST) between known Glutamin-(asparagin-)ase encoding genes and genes in *Microcoleus*. For a discussion of enzyme’s expressing Glutamate dehydrogenase (EC 1.4.1.2, 1.4.1.3, 1.4.1.4) look to the annotation of gene OID: 2505169473.

In addition to genes not predicted to IMG, literature suggested regulatory proteins such as PipX, a PII interaction protein, [[12]](http://198.128.29.183/genome/converter.php#ref-12)and glutamine synthase inactivating factors IF7 & IF17 [[12]](http://198.128.29.183/genome/converter.php#ref-12) found in *Synechocystis sp.* PCC 6803 were annotated as possible genes in *Microcoleus*.

**Nitrogen regulatory protein, PipX (PII interaction protein X)**

Gene OID: 2505167809

This is a probable PipX protein encoding gene inferred from homology and binding site conservation. Strong homologs were found in PDB (Q8YZH5, 62% identity, E-value 8.01E-29) [[21]](http://198.128.29.183/genome/converter.php#ref-21).

**Glutamine synthase inactivating factors IF7 & IF17**

Gene OID: 2505168417

This is a possible IF7 encoding gene inferred from homology with a known IF7 protein ( genbank: BAA17150, 72% identity, E-value 4.00E-21) [[12]](http://198.128.29.183/genome/converter.php#ref-12) around the same length (65-68 aa). Because IF7 is only a suspected posttranscriptional inhibitor of GSI nitrogen metabolism and not well studied further research is necessary to verify if this protein is in fact a GSI inhibitor.

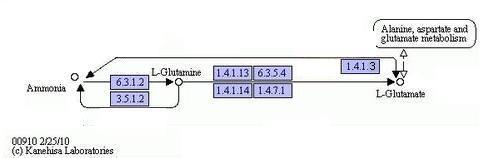
Gene OID: 2505168586

This is a possible IF17 encoding gene inferred from weak homology with a verified IF17 protein (genbank: BAA16994, 27% identity, E-value 1.00E-05) [[12]](http://198.128.29.183/genome/converter.php#ref-12). Because IF17 is only a suspected posttranscriptional inhibitor of GSI nitrogen metabolism and is not well studied, this requires further research to verify.

**Functional Studies**

Functional studies on genes suspected to encode the IF7 & IF17 inhibitors and Asparagine synthase (glutamine-hydrolysing) may be necessary in the future because their function is relatively uncertain due to weak homology. In a past study [[14]](http://198.128.29.183/genome/converter.php#ref-14) the glutamine synthetase of the cyanobacterium *Synechocystis sp.* PCC 6803 could be inactivated in vivo by ammonium addition, and the presence of the inhibitors were found by construction of a Synechocystis strain harboring a histidine-tagged modiﬁed version of GSI, allowing the puriﬁcation of the inactive enzyme and identification of inhibiting polypeptides. A similar study could be done on *Microcoleus*. Asparagine synthase function could be tested by constructing mutant strains of Microcoleus lacking the possible genes and observing what affect that has on growth in the presence of asparagine, which was done in a study on *Bacillus subtilis* [[19]](http://198.128.29.183/genome/converter.php#ref-19).

**Final Annotation**



S Fig. 3 Final annotated pathway map for *Microcoleus Vaginatus* PCC 9802 derived from IMG/ER predictions and other genome comparisons. Displays predicted enzymes.

Genes encoding ammonium transport proteins, enzymes involved in glutamiine and glutamate synthesis like GSI, GSII, and GDH, and nitrogen regulatory proteins like NtcA and PII/PipX (which are referenced throughout nitrogen metabolic pathways) are predicted and annotated within *Microcoleus vaginatus* PCC 9802. *Microcoleus* most likely assimilates ammonia as a nitrogen source.

**Discussion**

In light of the fact that not much research has been found about *Microcoleus’s* nitrogen capabilities, this gene annotation may provide insight into how it the cyanobacteria survives in the desert. For example, GDH is not strictly conserved in cyanobacteria, but because it does not require ATP it may be a more efficient pathway for glutamate synthesis when energy is scarce. In addition, GDH has been proven to be very thermostable in some bacteria [[3]](http://198.128.29.183/genome/converter.php#ref-3), a property that would be useful in the desert. GSIII, which is also not strictly conserved, has also been suggested to be more efficient than GSI in nitrogen deficient situations [[11]](http://198.128.29.183/genome/converter.php#ref-11), which could be another adaptive property in the desert, and perhaps during dormant periods. In the end, this annotation also serves as one of the key building blocks to constructing a clear picture of how *Microcoleus* metabolizes nitrogen.

**References**

[1] Belitsky BR, Wray LV Jr, Fisher SH, Bohannon DE, Sonenshein AL. "Role of TnrA in nitrogen source-dependent repression of Bacillus subtilis glutamate synthase gene expression." J Bacteriol. 2000 Nov; 182(21):5939-47.

[2] Binda C, Bossi RT, Wakatsuki S, Arzt S, Coda A, Curti B, Vanoni MA, Mattevi A. "Cross-talk and ammonia channeling between active centers in the unexpected domain arrangement of glutamate synthase." Structure. Dec 15, 2000; 8(12):1299-308.

[3] Britton KL, Yip KS, Sedelnikova SE, Stillman TJ, Adams MW, Ma K, Maeder DL, Robb FT, Tolliday N, Vetriani C, Rice DW, Baker PJ. "Structure determination of the glutamate dehydrogenase from the hyperthermophile Thermococcus litoralis and its comparison with that from Pyrococcus furiosus." J Mol Biol. Nov 12, 1999; 293(5):1121-32.

[4] Brown G, Singer A, Proudfoot M, Skarina T, Kim Y, Chang C, Dementieva I, Kuznetsova E, Gonzalez CF, Joachimiak A, Savchenko A, Yakunin AF. "Functional and structural characterization of four glutaminases from Escherichia coli and Bacillus subtilis." Biochemistry. May 27, 2008; 47(21):5724-35.

[5] Fokina O, Chellamuthu VR, Forchhammer K, Zeth K. "Mechanism of 2-oxoglutarate signaling by the Synechococcus elongatus PII signal transduction protein." Proc Natl Acad Sci U S A. Nov 16, 2010; 107(46):19760-5.

[6] Gotthardt D, Blancheteau V, Bosserhoff A, Ruppert T, Delorenzi M, Soldati T. "Proteomics fingerprinting of phagosome maturation and evidence for the role of a Galpha during uptake." Mol Cell Proteomics. 2006 Dec; 5(12):2228-43.

[7] Hanson TE, Forchhammer K, Tandeau de Marsac N, Meeks JC. "Characterization of the glnB gene product of Nostoc punctiforme strain ATCC 29133: glnB or the PII protein may be essential." Microbiology. 1998 Jun; 144 ( Pt 6)():1537-47.

[8] Knapp S, de Vos WM, Rice D, Ladenstein R. "Crystal structure of glutamate dehydrogenase from the hyperthermophilic eubacterium Thermotoga maritima at 3.0 A resolution." J Mol Biol. Apr 11, 1997; 267(4):916-32.

[9] Krajewski WW, Jones TA, Mowbray SL. "Structure of Mycobacterium tuberculosis glutamine synthetase in complex with a transition-state mimic provides functional insights." Proc Natl Acad Sci U S A. Jul 26, 2005; 102(30):10499-504.

[10] Montesinos ML, Muro-Pastor AM, Herrero A, Flores E. "Ammonium/methylammonium permeases of a Cyanobacterium. Identification and analysis of three nitrogen-regulated amt genes in synechocystis sp. PCC 6803." J Biol Chem. Nov 20, 1998; 273(47):31463-70.

[11] Muro-Pastor MI, Reyes JC, Florencio FJ. "Ammonium assimilation in cyanobacteria." Photosynth Res. 2005; 83(2):135-50.

[12] Mérida A, Candau P, Florencio FJ. "In vitro reactivation of in vivo ammonium-inactivated glutamine synthetase from Synechocystis sp. PCC 6803." Biochem Biophys Res Commun. Dec 16, 1991; 181(2):780-6.

[13] Nakasako M, Fujisawa T, Adachi S, Kudo T, Higuchi S. "Large-scale domain movements and hydration structure changes in the active-site cleft of unligated glutamate dehydrogenase from Thermococcus profundus studied by cryogenic X-ray crystal structure analysis and small-angle X-ray scattering." Biochemistry. Mar 13, 2001; 40(10):3069-79.

[14] Reyes JC, Florencio FJ. "A novel mechanism of glutamine synthetase inactivation by ammonium in the cyanobacterium Synechocystis sp. PCC 6803. Involvement of an inactivating protein." FEBS Lett. Jun 19, 1995; 367(1):45-8.

[15] Reyes JC, Florencio FJ. "Electron transport controls transcription of the glutamine synthetase gene (glnA) from the cyanobacterium Synechocystis sp. PCC 6803." Plant Mol Biol. 1995 Feb; 27(4):789-99.

[16] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes." Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[17] Wei TF, Ramasubramanian TS, Pu F, Golden JW. "Anabaena sp. strain PCC 7120 bifA gene encoding a sequence-specific DNA-binding protein cloned by in vivo transcriptional interference selection." J Bacteriol. 1993 Jul; 175(13):4025-35.

[18] Wray LV Jr, Atkinson MR, Fisher SH. "The nitrogen-regulated Bacillus subtilis nrgAB operon encodes a membrane protein and a protein highly similar to the Escherichia coli glnB-encoded PII protein." J Bacteriol. 1994 Jan; 176(1):108-14.

[19] Yoshida K, Fujita Y, Ehrlich SD. "Three asparagine synthetase genes of Bacillus subtilis." J Bacteriol. 1999 Oct; 181(19):6081-91.

[20] Zhao MX, Jiang YL, He YX, Chen YF, Teng YB, Chen Y, Zhang CC, Zhou CZ. "Structural basis for the allosteric control of the global transcription factor NtcA by the nitrogen starvation signal 2-oxoglutarate." Proc Natl Acad Sci U S A. Jul 13, 2010; 107(28):12487-92.

[21] Zhao MX, Jiang YL, Xu BY, Chen Y, Zhang CC, Zhou CZ. "Crystal structure of the cyanobacterial signal transduction protein PII in complex with PipX." J Mol Biol. Sep 24, 2010; 402(3):552-9.

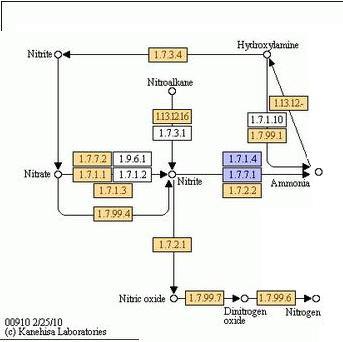
[22] van den Heuvel RH, Ferrari D, Bossi RT, Ravasio S, Curti B, Vanoni MA, Florencio FJ, Mattevi A. "Structural studies on the synchronization of catalytic centers in glutamate synthase." J Biol Chem. Jul 5, 2002; 277(27):24579-83.

## Nitrate Assimilation

Although atmospheric nitrogen (N2) is most likely not fixed by *M. vaginatus*, the cyanobacteria still metabolizes nitrogen from other nitrogen sources, such as ammonia, for its necessary functions. Nitrate may be one of those key nitrogen sources.

“Nitrate uptake and reduction to nitrite and ammonium are driven in cyanobacteria by photosynthetically generated assimilatory power, i.e., ATP and reduced ferredoxin. High-affnity nitrate and nitrite uptake takes place in different cyanobacteria through either an ABC-type transporter or a permease from the major facilitator superfamily (MFS). Nitrate reductase and nitrite reductase are ferredoxin-dependent metalloenzymes that carry as prosthetic groups a [4Fe–4S] center and Mo-bis-molybdopterin guanine dinucleotide (nitrate reductase) and [4Fe–4S] and siroheme centers (nitrite reductase). Nitrate assimilation genes are commonly found forming an operon with the structure: nir (nitrite reductase)-permease gene(s)-narB (nitrate reductase). When the cells perceive a high C to N ratio, this operon is transcribed from a complex promoter that includes binding sites for NtcA, a global nitrogen-control regulator that belongs to the CAP family of bacterial transcription factors, and NtcB, a pathway-speciﬁc regulator that belongs to the LysR family of bacterial transcription factors. Transcription is also affected by other factors such as CnaT, a putative glycosyl transferase, and the signal transduction protein PII. The latter is also a key factor for regulation of the activity of the ABC-type nitrate/nitrite transporter, which is inhibited when the cells are incubated in the presence of ammonium or in the absence of CO2.” [[3]](http://198.128.31.240/genome/converter.php#ref-3)

**Automated Annotation**



S Fig 1. Kegg Illustration of Nitrate Assimilation pathway of Nitrogen Metabolism. IMG/ER automated predictions are indicated by color: blue indicates a gene prediction in *M. vaginatus*, orange indicates a gene prediction in other genomes but not in *Microcoleus,* and white indicates no prediction was made. The original pathway map can be found at <http://www.genome.jp/kegg/pathway/map/map00910.html>. Does not include transport proteins.

**Nitrate transport ATP-binding subunits A,B,C, D (ABC Type)**

Gene OID: 2505171114, 2505171115, 2505171117, 2505171118 EC: 3.6.3.26

These genes are predicted to code for nitrate transport ATP-binding subunits due to strong homology with experimentally verified nitrate transport enzymes and a conserved gene neighborhood/cluster pattern with nitrate assimilation genes of *Synecchocystis sp.* [[10]](http://198.128.31.240/genome/converter.php#ref-10). However, IMG/ER falsely annotated these genes as different subunits of Nitrate transporters or different types.

Gene 2505171118 is predicted to code for subunit A; a strong homolog verified to code for nrtA was found in Swissprot (Q44292, 62% identity, E-value 0). [[5]](http://198.128.31.240/genome/converter.php#ref-5)

Gene 2505171117 is predicted to code for subunit B; a strong homolog verified to code for nrtB was found in Swissprot (Q51881, 63% identity, E-value 4.00E-119) [[8]](http://198.128.31.240/genome/converter.php#ref-8).

Gene 2505171115 is predicted to code for subunit C; a strong homolog verified to code for nrtC was found in Swissprot (P38045, 52% identity, E-value 0) [[11]](http://198.128.31.240/genome/converter.php#ref-11).

Gene 2505171114 is predicted to code for subunit D; a strong homolog verified to code for nrtD was found in Swissprot (P38046, 67% identity, E-value 1.00E-140) [[11]](http://198.128.31.240/genome/converter.php#ref-11).

**Nitrite reductase [NAD(P)H] large and small subunit**

Gene OID: 2505167495 &2505167494 EC: 1.7.1.4

Gene 2505167495 is predicted to be the large subunit of nitrite reductase [NAD(P)H] due to homology with experimentally verified nitrite reductase. Strong nitrite reductase homologs were not found in PDB for either subunit (possibly due to difficulty in crystallization) and the active sites of ferredoxin reductase homologs were not conserved. Despite this, nitrite reductase homologs were found in Swissprot (Uniprot: Q06458, 58% identity, E-value 0) [[7]](http://198.128.31.240/genome/converter.php#ref-7) and the gene clustered in the same neighborhood with the predicted small subunit.

The small subunit is predicted to be coded by gene 2505167494 due to homology with experimentally verified small subunits of nitrite reductase [NAD(P)H]l. Similar to the large subunit, the small subunit possessed strong, verified homologs in Swissprot (Uniprot: P0A229, 46% identity, E-value 3E-29) [[15]](http://198.128.31.240/genome/converter.php#ref-15) but not in PDB, most likely for the same reasons.

**Ferredoxin--nitrite reductase**

Gene OID: 2505171119 EC 1.7.7.1

This gene is predicted to code for Ferredoxin--nitrite reductase due to strong homology with experimentally verified NiR genes in cyanobacteria. The strongest homolog was found in Swissprot (Q51879, 75% identity , E-value 0) [[9]](http://198.128.31.240/genome/converter.php#ref-9). Although there is no active site conservation, similar to the situation with gene 2505167963, the strong homology found with cyanobacterial NiR and the conserved neighborhood with known nitrate assimilation gene clusters [[10]](http://198.128.31.240/genome/converter.php#ref-10) provide enough evidence to make this prediction.

**Global nitrogen regulator NtcA**

Gene OID: 2505167256

This is a probable global nitrogen regulator NtcA inferred from homology with experimentally verified homologs found in NCBI (P0A4U6, 87% identity, E-value 1.00E-141) [[14]](http://198.128.31.240/genome/converter.php#ref-14). In addition to this homology, 2-oxoglutarate binding site conservation was observed with another homology in PDB (P0A4U6) [[16]](http://198.128.31.240/genome/converter.php#ref-16).

**Nitrogen regulatory protein PII**

Gene OID: 2505170043

This is a probable PII encoding gene inferred from homology and binding site (with 2-oxoglutarate) conservation. Strong homologs were found in PDB (O30794, 92% identity, E-value 2.00E-69) [[6]](http://198.128.31.240/genome/converter.php#ref-6) as well as known binding sites (P0A3F4) [[4]](http://198.128.31.240/genome/converter.php#ref-4), which were conserved in this predicted protein.

**False Annotations**

Gene OID 2505167963

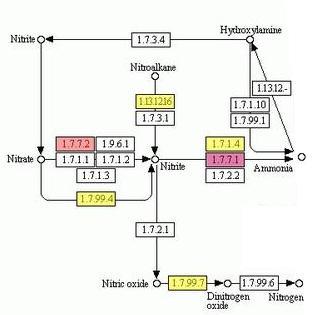
IMG automatically predicted this gene to code for Precorrin-3B synthase and EC 1.7.7.1--a contradiction most likely resulting from the use of pfam evidence for EC# assignment. In addition, the gene was in the neighborhood of predicted precorrin isomerase and methyltransferase, as opposed to the expected nitrate transporters and nitrate reductases [[10]](http://198.128.31.240/genome/converter.php#ref-10). A BLAST search against Swissprot yields mostly nitrite and sulfite reductases (Uniprot: P39661, 30% identity, E-value 3.00E-59) [[1]](http://198.128.31.240/genome/converter.php#ref-1)--this can be attributed to close homology with NiR.

Because no active sites for cyanobacterial Ferredoxin--nitrite reductase are present in literature, active site conservation is not observed among other bacterial/plant reductases, and a T-coffee multiple alignment using 250516793 and other NiR showed extremeley weak homology at the C and N terminals, the prediction that this enzyme codes for nitrite reductase behavior is most likely false.

Gene OIDs 2505169175, 2505169176, 2505169177, 2505169178, 2505169179:

Although these genes were predicted by IMG to code for nitrate transport proteins, they are probably bicarbonate-binding protein subunits due to strong homology with verified Bicarbonate-binding proteins. For example, 2505169179 has a strong homolog in Swissprot verified as Bicarbonate binding protein CmpA (Q55460, 64% identity, E-value 0) [[12]](http://198.128.31.240/genome/converter.php#ref-12). In addition, this neighborhood of genes deviates from the gene neighborhood/cluster pattern that nitrate assimilation genes are known to arrange in [[10]](http://198.128.31.240/genome/converter.php#ref-10).

**Other Genomes**



S Fig 2. IMG/ER illustration of enzyme conservation among 68 different cyanobacterial genomes in Kegg pathway.

Because many genes/enzymes were not predicted by IMG/ER to exist in *Microcoleus vaginatus*, known enzymes found in Uniprot, Swissprot, and NCBI were BLAST searched against the*Microcoleus* genome in order to find any homology that IMG could not detect. Unfortunately, many enzymes retrieved very weak or no homologs in *Microcoleus*(1.13.12.16, 1.7.3.1, 1.7.3.4, 1.7.99.7, 1.7.99.6, 1.7.99.1, 1.7.2.2, 1.7.1.10, 1.7.2.5, 1.13.12.-, and 1.9.6.1). Annotations below are the results of positive BLAST searches.

**Cytochrome c nitrite reductase**

EC 1.7.2.2

There is no gene prediction in *M. vaginatus* for Cytochrome c nitrite reductase. Cyanobacteria seem to not possess the Formate-dependent cytochrome c nitrite reductase, c552 subunit protein family, literature concerning cyanobacterial cytochrome c NiR does not exist, and BLAST searches with an E-value 1 threshold with known cytochrome c NiR (P0ABK9) [[2]](http://198.128.31.240/genome/converter.php#ref-2) against the *Microcoleus vaginatus* genome retrieve no homologs.

**Nitrate assimilation regulator ntcB**

Gene OID: 2505168074

This is a possible NtcB encoding gene due to strong homology with other predicted NtcB genes. The top experimentally verified homolog was found in Uniprot (P52693, 46% identity, E-value 4.00E-76) [[13]](http://198.128.31.240/genome/converter.php#ref-13); however, because very few experimentally verified NtcB homologs to base a decision on exist, more research is necessary to verify that this gene does encode NtcB.

**Nitrogen regulatory protein, PipX (PII interaction protein X)**

Gene OID: 2505167809

This is a probable PipX protein encoding gene inferred from homology and binding site conservation. Strong homologs were found in PDB (Q8YZH5, 62% identity, E-value 8.01E-29) [[17]](http://198.128.31.240/genome/converter.php#ref-17).

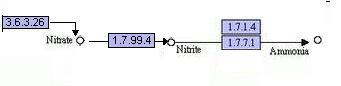
**Nitrate reductase**

Gene OID: 2505171113 EC: 1.7.99.4

Although IMG/ER does not predict this gene to exist in *Microcoleus* a gene was found probably to code for nitrate reductase due to strong homology with expermentally verified nitrate reductases, close active site conservation, and a conserved gene neighborhood region with known *Synechococcus elongatus* nitrate assimilation genes [[10]](http://198.128.31.240/genome/converter.php#ref-10). Although IMG/ER’ did not automatically predict NR to exist in*Microcoleus vaginatus* a BLAST against M. vaginatus with a NR gene from Synechococcus from Swissprot (P39458) [[11]](http://198.128.31.240/genome/converter.php#ref-11) yielded a strong homolog (63% identity, E-value 0). An additional BLAST against PDB revealed another strong homolog (P81186, 55% identity, E-value 4.71927E-135) whose active sites were closely matched (mismatches included Y713-L F689-L S615-T T422-S out of 13 active sites).

If this enzyme does have EC 1.7.99.4, it can use iron, molybdenum, and cytochrome cofactors as acceptors in the reduction, such that it encompasses functions 1.7.7.1/1.7.1.1/1.7.1.2/1.9.6.1.

**Final Annotation**



S Fig 3. Final Annotation diagram derived from KEGG illustration. Proteins without EC numbers not shown.

Probable genes coding for the assimilation of nitrate were found in this annotation. Although a large amount of functions went without prediction, *Microcoleus* most likely assimilates nitrate due to the presence of nitrogen regulator proteins, transport proteins, nitrate reductase and nitrite reductase,

**Functional Studies**

Out of the genes annotated above, the predicted nitrate regulatory protein encoding gene (2505168074) ntcB should be further researched in the laboratory to not only clarify this protein’s function but also better understand *Microcoleus’s* nitrate assimilation. This can be done by creating ntcB mutants and comparing its growth in nitrate and ammonia to nonmutated strains.[[13]](http://198.128.31.240/genome/converter.php#ref-13)

**Discussion**

In light of the fact that not much research has been found about *Microcoleus’s* nitrogen capabilities, this gene annotation may provide insight into how it the cyanobacteria survives in the desert. For example, by possessing EC 1.7.99.4 it is able to reduce nitrate with a variety of cofactors--this may aid its existence in a variety of environments. In addition, its seeming lack of genes that allow it to use hydroxlyamine, nitroalkane, or produce nitrogen gas may be due to a lack of information on its nitrogen metabolism or a different, unknown pathway that could exist inside *M. vaginatus*.

**References**

[1] "" . ; ():.

[2] Darwin A, Hussain H, Griffiths L, Grove J, Sambongi Y, Busby S, Cole J. "Regulation and sequence of the structural gene for cytochrome c552 from Escherichia coli: not a hexahaem but a 50 kDa tetrahaem nitrite reductase." Mol Microbiol. 1993 Sep; 9(6):1255-65.

[3] Flores E, Frías JE, Rubio LM, Herrero A. "Photosynthetic nitrate assimilation in cyanobacteria." Photosynth Res. 2005; 83(2):117-33.

[4] Fokina O, Chellamuthu VR, Forchhammer K, Zeth K. "Mechanism of 2-oxoglutarate signaling by the Synechococcus elongatus PII signal transduction protein." Proc Natl Acad Sci U S A. Nov 16, 2010; 107(46):19760-5.

[5] Frías JE, Flores E, Herrero A. "Nitrate assimilation gene cluster from the heterocyst-forming cyanobacterium Anabaena sp. strain PCC 7120." J Bacteriol. 1997 Jan; 179(2):477-86.

[6] Hanson TE, Forchhammer K, Tandeau de Marsac N, Meeks JC. "Characterization of the glnB gene product of Nostoc punctiforme strain ATCC 29133: glnB or the PII protein may be essential." Microbiology. 1998 Jun; 144 ( Pt 6)():1537-47.

[7] Lin JT, Goldman BS, Stewart V. "Structures of genes nasA and nasB, encoding assimilatory nitrate and nitrite reductases in Klebsiella pneumoniae M5al." J Bacteriol. 1993 Apr; 175(8):2370-8.

[8] Merchán F, Kindle KL, Llama MJ, Serra JL, Fernández E. "Cloning and sequencing of the nitrate transport system from the thermophilic, filamentous cyanobacterium Phormidium laminosum: comparative analysis with the homologous system from Synechococcus sp. PCC 7942." Plant Mol Biol. 1995 Jul; 28(4):759-66.

[9] Merchán F, Prieto R, Kindle KL, Llama MJ, Serra JL, Fernández E. "Isolation, sequence and expression in Escherichia coli of the nitrite reductase gene from the filamentous, thermophilic cyanobacterium Phormidium laminosum." Plant Mol Biol. 1995 Mar; 27(5):1037-42.

[10] Ohashi Y, Shi W, Takatani N, Aichi M, Maeda S, Watanabe S, Yoshikawa H, Omata T. "Regulation of nitrate assimilation in cyanobacteria." J Exp Bot. 2011 Feb; 62(4):1411-24.

[11] Omata T, Andriesse X, Hirano A. "Identification and characterization of a gene cluster involved in nitrate transport in the cyanobacterium Synechococcus sp. PCC7942." Mol Gen Genet. 1993 Jan; 236(2-3):193-202.

[12] Omata T, Gohta S, Takahashi Y, Harano Y, Maeda S. "Involvement of a CbbR homolog in low CO2-induced activation of the bicarbonate transporter operon in cyanobacteria." J Bacteriol. 2001 Mar; 183(6):1891-8.

[13] Suzuki I, Horie N, Sugiyama T, Omata T. "Identification and characterization of two nitrogen-regulated genes of the cyanobacterium Synechococcus sp. strain PCC7942 required for maximum efficiency of nitrogen assimilation." J Bacteriol. 1995 Jan; 177(2):290-6.

[14] Wei TF, Ramasubramanian TS, Pu F, Golden JW. "Anabaena sp. strain PCC 7120 bifA gene encoding a sequence-specific DNA-binding protein cloned by in vivo transcriptional interference selection." J Bacteriol. 1993 Jul; 175(13):4025-35.

[15] Wu JY, Siegel LM, Kredich NM. "High-level expression of Escherichia coli NADPH-sulfite reductase: requirement for a cloned cysG plasmid to overcome limiting siroheme cofactor." J Bacteriol. 1991 Jan; 173(1):325-33.

[16] Zhao MX, Jiang YL, He YX, Chen YF, Teng YB, Chen Y, Zhang CC, Zhou CZ. "Structural basis for the allosteric control of the global transcription factor NtcA by the nitrogen starvation signal 2-oxoglutarate." Proc Natl Acad Sci U S A. Jul 13, 2010; 107(28):12487-92.

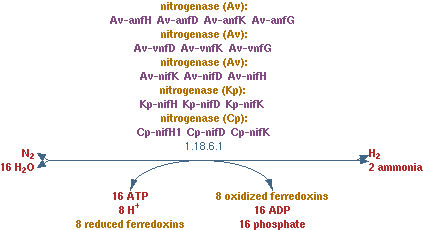
[17] Zhao MX, Jiang YL, Xu BY, Chen Y, Zhang CC, Zhou CZ. "Crystal structure of the cyanobacterial signal transduction protein PII in complex with PipX." J Mol Biol. Sep 24, 2010; 402(3):552-9.

## Nitrogen Fixation

Nitrogen is a common building block of amino and nucleic acids, essential to life on Earth. Despite the fact that about 80% of the atmosphere is N2, it is a limiting nutrient in many ecosystems. This is because the dinitrogen molecule can not be used directly for biosynthesis and must be converted to a reduced form to become bioavailable. The process of the reductive breakage of the nitrogen-nitrogen bond forming two ammonia molecules is called nitrogen fixation, and can be performed biologically only by certain bacteria and archaea that possess the nitrogenase enzyme system. The nitrogenase complex is extremely sensitive to molecular oxygen and is only active under anaerobic conditions.

Nitrogen fixation is accompanied by an obligatory reduction of protons to H2. Of the diazotrophic microbes known to possess this function, some species of *Cyanobacteria* are among the best categorized. While cyanobacteria are photoautotrophic microorganisms, light requirement for H2 generation using the nitrogenase complex varies among different cyanobacterial species. *Microcoleus vaginatus* is not predicted to have the enzyme necessary for nitrogen fixation, but an annotation is done anyways just to check.

**Automated Annotation**



S Fig. 1: Nitrogen fixation pathway requires the nitrogenase enzyme (EC 1.18.6.1)

**Nitrogenase**

Gene OID: 2505166005 EC: 1.18.6.1

A search was done in IMG/ER for any genes in *Microcoleus vaginatus* predicted to code for EC 1.18.6.1. This search gave no hits. A UniProt search was then done for EC 1.18.6.1, and the FASTA sequences were narrowed down to those for reviewed cyanobacteria. These FASTA sequences were BLASTed against the *Microcoleus vaginatus* genome in IMG/ER. Gene 2505166005 kept on showing up in many of these searches. This gene was predicted by IMG/ER to be a light-independent protochlorophyllide reductase, iron-sulfur ATP-binding protein. A NCBI BLAST search [1] (NCBI: Q8YM62) gave as a top hit another light-independent protochlorophyllide reductase, iron-sulfur ATP-binding protein, with 93% identity and an e-value of 1E-147. A reverse BLAST then gave gene 2505166005 as the first result.

A search was also done with PDB to double-check. The PDB BLAST search gave as the first hit an article on the crystal structure of a light-independent protochlorophyllide reductase protein, but also mentioned how this protein is a homologue of the nitrogenase Fe protein [3] (PDB: 3FWY). A reverse BLAST search gave gene 2505166005 as the first result. The second hit was an article on the nitrogenase Fe protein [2] (PDB: 1FP6). A reverse BLAST search gave gene 2505166005 as the top result. This article mentioned different active site residues for the nitrogenase protein. These active site residues were checked against the *Microcoleus vaginatus* gene. 14 out of 18 active sites matched. The mismatches were D214 to I214, Q218 to R218, E221 to R 221 and Q236 to Y236. The purpose of these sites was searched for in the article. D214 is needed for its hydrogen bonding carbonyl oxygen, which are not present in I214. E221 is needed for its hydrogen bonding side chain oxygen, which are not present in R221. Q218 and Q236 are needed for their side chain interactions with the adenine base bound to the Fe protein, through their side chain oxygens. R218 and Y236 do not have side chain oxygens. These results suggest that gene 2505166005 does not code for a nitrogenase protein.

As a final check, a multiple sequence alignment was done on T-Coffee [4]. The FASTA sequences used for comparison were reviewed cyanobacteria sequences from UniProt. This showed that at site 214, most cyanobacteria had D214, and no other cyanobacteria had I214. At site 218, no other cyanobacteria had R218 and most of them had Q218. At site 221, no other cyanobacteria had R221, and most of them had E221. Finally, at site 236, no other cyanobacteria had Y236, and most of them had Q236. From this analysis and the active site residues, it seems that gene 2505166005 does not code for a nitrogenase protein.

**Discussion**

*Microcoleus vaginatus* is not predicted to undergo nitrogen fixation, because it is missing the essential nitrogenase enzyme (EC 1.18.6.1). However, an experiment should still be done in the lab to see if *Microcoleus vaginatus* can fix its own nitrogen to make sure. From the annotation though, it seems that *Microcoleus vaginatus* is one of the cyanobacteria that does not undergo nitrogen fixation.

**References**

[1] Kaneko,T., Nakamura,Y., Wolk,C.P., et al: Complete genomic sequence of the filamentous

nitrogen-fixing cyanobacterium Anabaena sp. strain PCC 7120. DNA Res. 8 (5), 205-213

(2001).

[2] Jang, S.B., Seefeldt, L.C., Peters, J.W.: Insights into nucleotide signal transduction in

nitrogenase: structure of an iron protein with MgADP bound. (2000) Biochemistry 39:

14745-14752

[3] Sarma, R., Barney, B.M., Hamilton, T.L.: Crystal Structure of the L Protein of Rhodobacter

sphaeroides Light-Independent Protochlorophyllide Reductase with MgADP Bound: A

Homologue of the Nitrogenase Fe Protein. (2008) Biochemistry 47: 13004-13015

[4] T-Coffee: A novel method for fast and accurate multiple sequence alignment. (2000) J Mol  
 Biol. 302(1):205-17.

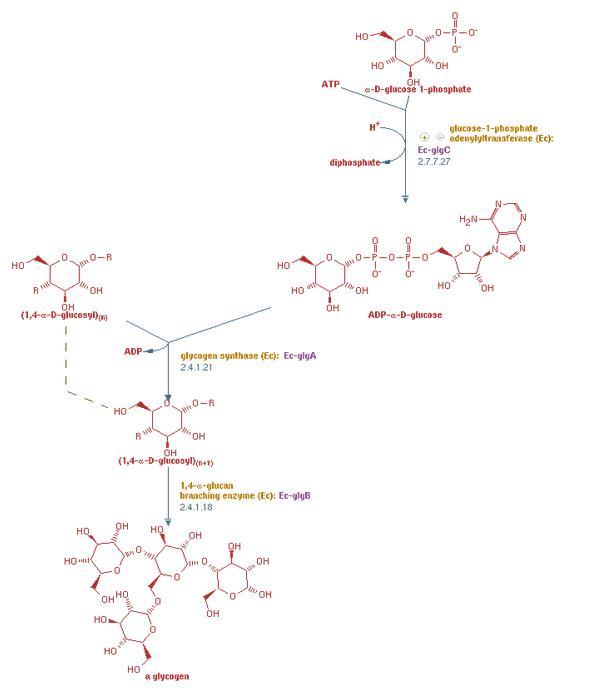
# Glycogen Biosynthesis

Glycogen and starch, megadalton-sized glucose polymers, are the major reservoir of readily available energy and carbon compounds in most living organisms, ranging from archaea, eubacteria and yeasts, up to higher eukaryotes including plants and animals. The linkages between the glucose units in glycogen are comprised of approximately 90% α-1,4 linkages, with the remainder being α-1,6 branch glucosyl linkages. The average chain length is about 12 to 14 glucose units.

A key enzyme in the pathway is glycogen (or starch) synthase. Glycogen and starch synthases are classified in two large and distinct famillies: the mammalian and yeast enzymes utilize UDP-D-glucose as the glucosyl group donor, while bacterial and plant enzymes prefer ADP-α-D-glucose.

In bacteria, ADP-glucose is transferred to either a maltodextrin or glycogen primer forming a new α-1,4-glucosidic linkage. The first enzyme of the pathway, glucose-1-phosphate adenylyltransferase, is rate-limiting. After chain elongation the formation of the branched α-1,6-glucosidic linkages is catalyzed by the glycogen branching enzyme from the growing polyglucose chain.

**Automated Annotation**



S Fig. 1: Metacyc illustrated pathway of glycogen biosynthesis pathway

**1,4-alpha-glucan branching enzyme**

Gene OIDs: 2505169365 EC: 2.4.1.18

There were 6 genes predicted by IMG/ER to code for this protein. The genes were analyzed in order, starting with gene 2505169365. For this gene, a NCBI BLAST search was done, and the top two hits were malto-oligosyltrehalose trehalohydrolases [[5]](http://198.128.31.59/genome/converter.php#ref-5) with e-values of 3E-38 and 6E-36 and 28% and 26% identity. A reverse BLAST hit gave another gene in IMG/ER predicted to be a malto-oligosyltrehalose trehalohydrolase. After these results though, most of the other results were 1,4-alpha-glucan branching enzymes, with e-values from 5E-30 to 3E-20 and around 25% to 30% identity. A reverse BLAST search gave one of the aforementioned genes as the firs hit, with this gene being the second hit. A PDB search was also done and gave as a hit a glycosyltrehalose trehalohydrolase [[6]](http://198.128.31.59/genome/converter.php#ref-6) with an e-value of 1.34664E-40 and 28% identity. Seven active site residues were mentioned, and 4 of them matched with the query protein. One mismatch was due to a deletion: there was no amino acid that lined up with the subject protein at that site. The other mismatches were C298L and R363Y. The C298 site was explained by the paper to be important for its sulfur. However, it says that this sulfur is only there for stability and structural integrity, not essential dimerization nor activity. The R363 site was not mentioned. Besides these active sites, there were also a list of the amino acids at the sites from 345 to 360, but there as a deletion along this stretch, such that there was no part of the query amino acid sequence that lined up to this stretch of the subject amino acid sequence. A reverse BLAST was also done and showed as the first hit another gene that is predicted by IMG/ER to code for malto-oligosyltrehalose trehalohydrolase, and then this gene as the second hit. Nonetheless, the data as a whole suggests that this gene might have been misannotated and may be a malto-oligosyltrehalose trehalohydrolase (EC 3.2.1.1) instead.

Gene OID: 2505169371 EC 2.4.1.18

Gene 2505169371 was predicted to not code for a protein with EC 2.4.1.18. Both NCBI {{NCBI: Q88FN1}) and BLAST searces [[12]](http://198.128.31.59/genome/converter.php#ref-12) gave weak hits. NCBI gave hits with e-values from 3.7 to 8.3 and 30% to 40% identity. PDB gave as hits published articles with e-values all higher than 4. In general, the BLAST hits for this gene were poor. It appears that the gene does not code for EC 2.4.1.18.

Gene OID: 2505169372 EC 2.4.1.18

NCBI and PDB BLAST searches done for this gene did not give other 1,4-alpha-glucan branching enzymes as top hits. The first NCBI hit for 1,4-alpha-glucan [[11]](http://198.128.31.59/genome/converter.php#ref-11) had an e-value of 0.73 and 44% identity, but the amino acid sequence is only 79 amino acids long. PDB did not give any hits for 1,4-alpha-glucan. Therefore it is predicted that this hypothetical protein is not 1,4-alpha-glucan, but rather maltogenic alpha-amylase (EC: 3.2.1.135), the top hits from the NCBI and PDB BLAST searches.

Gene OID: 2505169373 EC 2.4.1.18

For this gene, a NCBI BLAST showed most of the hits as being for 1,4-alpha-glucan-branching enzymes, but the top hit and some of the other hits were for malto-oligosyltrehalose trehalohydrolase. The top hit for malto-oligosyltrehalose trehalohydrolase [[5]](http://198.128.31.59/genome/converter.php#ref-5) had an e-value of 9E-11 and 48% identity, while the next best hit for 1,4-alpha-glucan-branching enzyme [[14]](http://198.128.31.59/genome/converter.php#ref-14) had an e-value of 5E-10 and 46% identity. The top hits from PDB were all for malto-oligosyltrehalose trehalohydrolases [[7]](http://198.128.31.59/genome/converter.php#ref-7). The top hit had an e-value of 1.94584E-11 and 48% identity. A reverse BLAST with the PDB FASTA seqeunce gave another gene in *Microcoleus vaginatus* that is predicted to code for malto-oligosyltrehalose trehalohydrolase as the top hit. This gene was a short one, with only 78 amino acids, so it is hard to say for certain if it is a 1,4-alpha-glucan-branching enzyme or trehalohydrolase. From the data though, it seems that this gene codes for the trehalohydrolase and not the branching enzyme.

Gene OID: 2505169374 EC 2.4.1.18

For the NCBI BLAST done with this gene, the top results were all for 1,4-alpha-glucan-branching enzymes [[8]](http://198.128.31.59/genome/converter.php#ref-8), with the top hit having an e-value of 0.021 and 38% identity. However, a reverse BLAST search showed gene 2505170155 as the top hit. There were also some hits to 5'-AMP-activated protein kinase subunit beta-1 [[13]](http://198.128.31.59/genome/converter.php#ref-13), with the best hit having an e-value of 0.11 and 38% identity. The top PDB BLAST hits were also to 5’-AMP-activated protein kinase subunit beta-1 [[10]](http://198.128.31.59/genome/converter.php#ref-10), with the top hit having an e-value of 0.00729827 and 40% identity. Further down the list was a hit to a 1,4-alpha-glucan-branching enzyme (PDB: 3K1D), with an e-value of 2.59577 and 35% identity. All these hits had high e-values, and the gene was short, only having 74 amino acids. A reverse BLAST hit with the PDB FASTA seqeunce gave no hits, so it seems that there are no genes that code for the 5’-AMP-activated protein kinase subunit beta-1. This suggests that *Microcoleus vaginatus* does not have a 5’-AMP-activated protein kinase. Therefore, it seems unlikely that the gene would code for the protein kinase subunit, so if it codes for a protein, it is probably for the 1,4-alpha-glucan-branching enzyme.

Gene OID: 2505170155 EC 2.4.1.18

A NCBI BLAST search gave other 1,4-alpha-glucan-branching enzymes as top hits [[9]](http://198.128.31.59/genome/converter.php#ref-9), with e-values around 0 and around 80% identity. A PDB BLAST search gave articles on the crystal structure of 1,4-alpha-glucan-branching enzyme as top hits [[2]](http://198.128.31.59/genome/converter.php#ref-2), with the top hit having an e-value of 0 and 51% identity. An active site residue analysis was attempted but was not successful because the active site numbering system in the article did not match with the PDB active site numbering in the multiple sequence alignment, such that if the article mentioned a site (say E201), the site on the multiple sequence alignment for the query protein was a different amino acid. Despite this, the numerical data and strong hits are good evidence that this gene indeed codes for the protein, along with the fact that it is around 740 amino acids long, which is the length of other cyanobacterial 1,4-alpha-glucan-branching enzymes as found in the UniProt database.

**Glucose-1-phosphate adenylyltransferase**

Gene OID: 2505167366 EC: 2.7.7.27

This gene is predicted to code for glucose-1-phosphate adenylyltransferase due to strong homology with experimentally verified enzymes with EC 2.7.7.27 and 100% active site conservation. Homologs were found in Swissprot (80% identity, E-value 0) [[1]](http://198.128.31.59/genome/converter.php#ref-1) and PDB (E-value 6.36689E-53, 53% identity) [[3]](http://198.128.31.59/genome/converter.php#ref-3).

**Glycogen synthases, ADP-glucose type**

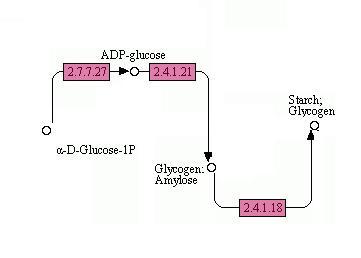
Gene OID: 2505169158 & 2505168172 EC: 2.4.1.21

These genes are predicted to code for glycogen synthase, ADP-glucose type due to strong homology with experimentally verified enzymes with EC 2.4.1.21 and close active site conservation.

For 2505169158, 8 out of the 12 mentioned active site residues matched while the mismatches are similar (S298-T, G353-L, I297-V, T381-S), and for 2505168172 10 out of 12 mentioned active site residuesmatched while the mismatches are similar (Y354-F, S298-G). In both cases, the mismatches were residues that contributed to weak hydrogen bonding or Van der Waals forces, so it is very possible that this protein still retains its effective catalytic function.

Homologs for 2505169158 and 2505168172 respectively were found in Swissprot (same hit, 79% identity, E-value 0) {{Q8YVU5}} and PDB ( for \*158 33% identity, E-value 1.20E-68; for \*172 30% identity, E-value 7.81E-52) [[4]](http://198.128.31.59/genome/converter.php#ref-4).

**Comparison with other Genomes**





S Fig. 2: Comparison of genes found in *Microcoleus vaginatus* with other cyanobacteria in IMG/ER for glycogen biosynthesis

All three of the enzymes needed for glycogen biosynthesis were predicted to be found in more than 75% of other cyanobacteria. There were no genes found to be specific to *Microcoleus vaginatus*, and there were no genes found to be in other cyanobacteria but not in *Microcoleus vaginatus*.

**Functional Studies**

Functional studies should be done on genes 2505169158 and 2505168172, which were both predicted to code for EC 2.4.1.21. A preliminary experiment should be done where both genes are repressed, and it should be seen if the protein is still expressed. If the protein is not expressed, then two more experiments should be done to see which gene is essential for the expression of EC 2.4.1.21, or if both genes are essential. In each experiment, one gene should be repressed and the organism should be observed to see if it still expresses a protein with EC 2.4.1.21 function.

Functional studies could also be done on genes 2505169365, 2505169371, 2505169372, 2505169373 and 2505169374. These genes were all short amino acid sequences that were predicted by IMG/ER to potentially code for 1,4-alpha-glucan-branching enzymes. Further annotation showed that most of these genes could possibly code for another enzyme, and that one of them was probably not a 1,4-alpha-glucan-branching enzyme. Further studies could be done to test these hypotheses. The amino acid sequence lengths suggest that these genes are not coding for 1,4-alpha-glucan-branching enzyme, because this enzyme is usually code for by around 700 amino acids. These genes are probably misannotated genes.

**Discussion**

All the enzymes needed for the pathway were found to have genes that coded for them, but two of the enzymes had multiple gene predictions. For EC 2.4.1.21, both genes 2505169158 and 2505168172 were predicted to code for the enzyme. It is interesting that this is the case. As mentioned before, studies should be done to see if both genes are necessary for the expression of the protein. If not, it would be interesting to do further studies to see if having both genes has some sort of evolutionary advantage, or if the two genes are somehow reserved for different circumstances that the organism might experience.

For EC 2.4.1.18, there were six predicted genes for the enzyme. 2505170155 is most likely the actual gene that codes for the enzyme. The other genes, 2505169365, 2505169371, 2505169372, 2505169373 and 2505169374, were probably short reads which were misannotated. Functional studies should be done to correct these wrong annotations.

**References**

[1] Abad MC, Binderup K, Rios-Steiner J, Arni RK, Preiss J, Geiger JH. "The X-ray

crystallographic structure of Escherichia coli branching enzyme." J Biol Chem. Nov 1,

2002; 277(44):42164-70.

[2] Abad MC, Binderup K, Rios-Steiner J, Arni RK, Preiss J, Geiger JH. "The X-ray

crystallographic structure of Escherichia coli branching enzyme." J Biol Chem. Nov 1,

2002; 277(44):42164-70.

[3] Cupp-Vickery JR, Igarashi RY, Perez M, Poland M, Meyer CR. "Structural analysis of

ADP-glucose pyrophosphorylase from the bacterium Agrobacterium tumefaciens."

Biochemistry. Apr 15, 2008; 47(15):4439-51.

[4] Cupp-Vickery JR, Igarashi RY, Perez M, Poland M, Meyer CR. "Structural analysis of

ADP-glucose pyrophosphorylase from the bacterium Agrobacterium tumefaciens."

Biochemistry. Apr 15, 2008; 47(15):4439-51.

[5] Feese MD, Kato Y, Tamada T, Kato M, Komeda T, Miura Y, Hirose M, Hondo K, Kobayashi K,

Kuroki R. "Crystal structure of glycosyltrehalose trehalohydrolase from the

hyperthermophilic archaeum Sulfolobus solfataricus." J Mol Biol. Aug 11, 2000;

301(2):451-64.

[6] Feese MD, Kato Y, Tamada T, Kato M, Komeda T, Miura Y, Hirose M, Hondo K, Kobayashi K,

Kuroki R. "Crystal structure of glycosyltrehalose trehalohydrolase from the

hyperthermophilic archaeum Sulfolobus solfataricus." J Mol Biol. Aug 11, 2000;

301(2):451-64.

[7] Feese MD, Kato Y, Tamada T, Kato M, Komeda T, Miura Y, Hirose M, Hondo K, Kobayashi K,

Kuroki R. "Crystal structure of glycosyltrehalose trehalohydrolase from the

hyperthermophilic archaeum Sulfolobus solfataricus." J Mol Biol. Aug 11, 2000;

301(2):451-64.

[8] Hoskins J, Alborn WE Jr, Arnold J, Blaszczak LC, Burgett S, DeHoff BS, Estrem ST, Fritz L,

Fu DJ, Fuller W, Geringer C, Gilmour R, Glass JS, Khoja H, Kraft AR, Lagace RE,

LeBlanc DJ, Lee LN, Lefkowitz EJ, Lu J, Matsushima P, McAhren SM, McHenney M,

McLeaster K, Mundy CW, Nicas TI, Norris FH, O'Gara M, Peery RB, Robertson GT,

Rockey P, Sun PM, Winkler ME, Yang Y, Young-Bellido M, Zhao G, Zook CA, Baltz RH,

Jaskunas SR, Rosteck PR Jr, Skatrud PL, Glass JI. "Genome of the bacterium

Streptococcus pneumoniae strain R6." J Bacteriol. 2001 Oct; 183(19):5709-17.

[9] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A,

Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[10] Polekhina G, Gupta A, van Denderen BJ, Feil SC, Kemp BE, Stapleton D, Parker MW.

"Structural basis for glycogen recognition by AMP-activated protein kinase." Structure.

2005 Oct; 13(10):1453-62.

[11] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[12] Simpson AA, Chandrasekar V, Hébert B, Sullivan GM, Rossmann MG, Parrish CR. "Host

range and variability of calcium binding by surface loops in the capsids of canine and

feline parvoviruses." J Mol Biol. Jul 14, 2000; 300(3):597-610.

[13] Stapleton D, Gao G, Michell BJ, Widmer J, Mitchelhill K, Teh T, House CM, Witters LA,

Kemp BE. "Mammalian 5'-AMP-activated protein kinase non-catalytic subunits are

homologs of proteins that interact with yeast Snf1 protein kinase." J Biol Chem. Nov 25,

1994; 269(47):29343-6.

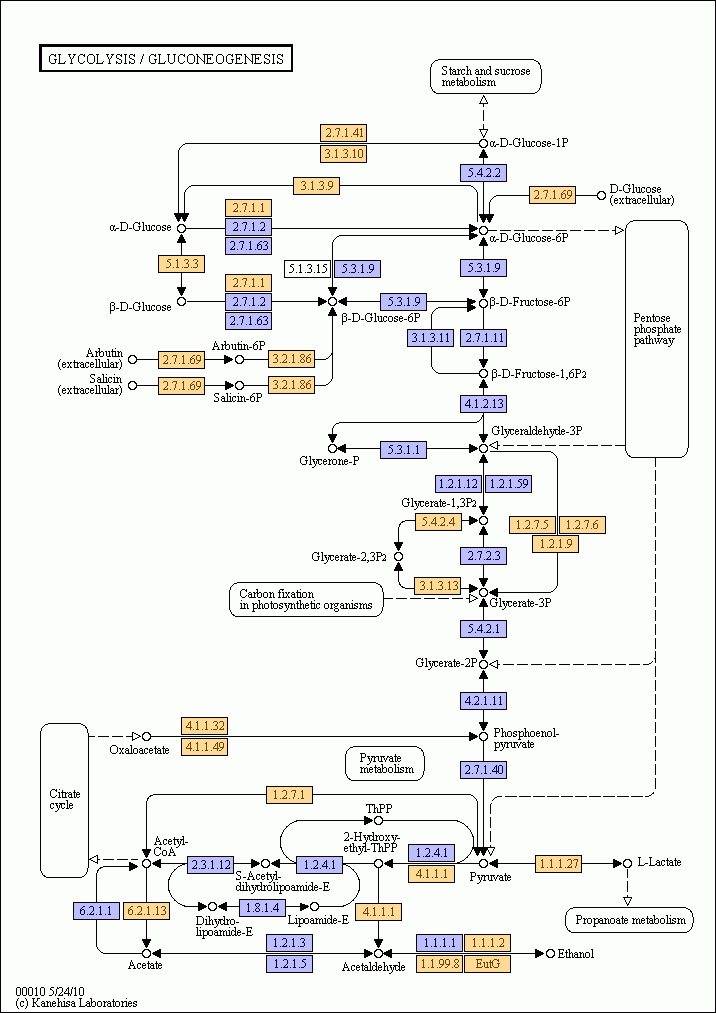
[14] Takami H, Takaki Y, Uchiyama I. "Genome sequence of Oceanobacillus iheyensis isolated

from the Iheya Ridge and its unexpected adaptive capabilities to extreme environments."

Nucleic Acids Res. Sep 15, 2002; 30(18):3927-35.

# Glycolysis

According to MetaCyc, glycolysis, which was first studied as a pathway for the utilization of glucose, is one of the components of central metabolism, the other two being the pentose phosphate pathway and the TCA cycle. As such, its functioning is essential under all conditions of growth because it produces six ( β-D-glucose-6-phosphate, D-fructose-6-phosphate, dihydroxyacetone phosphate 3-phospho-D-glycerate, phosphoenolpyruvate, and pyruvate) of the 13 precursor metabolites that are the starting materials for the biosynthesis of building blocks for macromolecules and other needed small molecules. Glycolysis can be found, if at least in part, in all organisms. Glycolysis has evolved to fulfill two essential functions: i) it oxidizes hexoses to generate ATP, reductants and pyruvate, and ii) being an amphibolic pathway (pathway that involves both catabolism and anabolism), it can reversibly produce hexoses from various low-molecular weight molecules. Because various degradation pathways feed into glycolysis at many different points, glycolysis or portions of it run in the forward or reverse direction, depending on the carbon source being utilized, in order to satisfy the cell's need for precursor metabolites and energy. This switching of direction is possible because all but two of the enzymatic reactions comprising glycolysis are reversible, and the conversions catalyzed by the two exceptions are rendered functionally reversible by other enzymes (fructose-1,6-bisphosphatase and phosphoenolpyruvate synthetase) that catalyze different irreversible reactions flowing in the opposite direction.

**Automated Annotation [from KEGG and MetaCyc]**

S Fig. 1: KEGG Automatic Annotation of Microcoleus Vaginatus Genome (blue indicates found genes, orange indicates missing genes)

MetaCyc illustration of pathway can be found at <http://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=GLYCOLYSIS&detail-level=2&detail-level=1&detail-level=2>

(image is too big to fit into report)

**6-Phosphofructokinase:**

Gene OID: 2505167140 EC: 2.7.1.11

Gene is confirmed to code for enzyme 2.7.1.11. A NCBI BLAST search [11] (NCBI: Q8YKG3) gave as top results other 6-phosphofructokinases. A PDB BLAST search [21] (PDB: 6PFK) gave an article that described active sites for the enzyme. These active site residues were compared to the ones for the enzyme coded for by the gene in *Microcoleus vaginatus*. 12 out of the 15 active site residues matched. The ones that weren’t conserved were R72, T158 and E161. These residues were N72, A158 and S161 for *Microcoleus vaginatus*.

To double-check these active sites, a multiple sequence alignment was done with T-Coffee [24] using FASTA sequences from UniProt [24] for comparison. This showed that for site 72, many other cyanobacteria had N72 instead of R72. At site 158, all but one cyanobacteria had A158 instead of T158. For site 161, some of the other cyanobacteria had S161 instead of E161. All these substitutions were still considered good matches by T-Coffee (as indicated by color), and the 6-phosphofructokinase genes from different cyanobacteria matched up well.

**3-phosphoglycerate kinase**

Gene OID 2505166234 EC 2.7.2.3

This gene has been confirmed to encode for phosphoglycerate kinase, the enzyme responsible for function 2.7.2.3. Strong homology between phosphoglycerate kinase in other annotated genomes in the Swissprot database [5] (NCBI:Q3MF40) and PDB [3] (PDB: 1VPE), including close active site matching, confirms this. Only one active site does not match, with D199 being replaced with S199, which only differ in an oxygen group on the side chain.

**Fructose-Bisphosphate Aldolase**

Gene OID: 2505166282 and 2505169777 EC: 4.1.2.13

Two hits were found for this corresponding enzyme. Gene 2505166282 was predicted to be a type II Fructose-Bisphosphate Aldolase. Gene 2505169777 was also predicted to be a Fructose-Bisphosphate Aldolase, and perhaps an archaeal one.

For gene 2505166282, it was predicted to code for a type II Fructose-Bisphosphate Aldolase (FBA). This was supported by a NCBI BLAST which gave as top results other FBAs. Additionally, a PDB BLAST [9] (PDB: 3GAC) gave an article with active site residues for class II FBA. These active site residues matched with the ones in microcoleus vaginatus, except for one at position 256, where S256 became T256. The article mentioned that S256 was important for its side chain hydroxyl group. T256 would also have a side chain hydroxyl group, and serine and threonine are very similar in structure. Finally, a multiple sequence alignment using FASTA sequences from UniProt showed that for cyanobacteria, the T256 active site residue is completely conserved.

For gene 2505169777, it was also predicted to be a FBA. A NCBI BLAST also gave as top results other FBAs. A PDB BLAST [15] (PDB: 1OJX) gave an article about archael FBAs. The active sites from this article matched with the active sites in microcoleus vaginatus, except for site 22, where A22 was P22 for microcoleus vaginatus. The article says however that A22 is not conserved, so the enzyme from microcoleus vaginatus could be similar to archael FBAs.

**Triose-phosphate isomerase**

Gene OID: 2505168240 EC: 5.3.1.1

The gene (2505168240) is confirmed to code for the triose-phosphate isomerase enzyme (EC 5.3.1.1). A NCBI BLAST search against the swissprot database gives other triose-phosphate isomerases as top hits. A PDB BLAST [8](PDB:1BTM ) search gave an article on the triose-phosophate isomerase enzyme from *Bacillus stearothermophilus*, which is a thermophilic bacteria, similar to *Microcoleus vaginatus* which can endure high desert temperatures. Additionally, all the active sites mentioned in this article matched with the ones in *Microcoleus vaginatus*.

**Glyceraldehyde-3-Phosphate Dehydrogenase Type 1**

Gene OID: 2505170636 EC: 1.2.1.12

Gene is confirmed to code for glyceraldehyde-3-phosphate dehydrogenase (GAPD). A NCBI BLAST search gave as top results other GAPDs. A PDB BLAST [27](PDB: 1DC3) search gave an article with active site residues for a GAPD, and NCBI BLAST provided very strong evidence for the same conclusion [17] (NCBI: P34916). The active site residues all matched with the ones in *Microcoleus vaginatus*.

**Pyruvate Kinase**

Gene OID: 2505167661 EC: 2.7.1.40

Gene is confirmed to code for pyruvate kinase. A NCBI BLAST returned a near-perfect match with another pyruvate kinase [12](NCBI: P73534), and gave as top results other pyruvate kinases. A PDB BLAST [18](PDB:1PKY) search gave an article on pyruvate kinase structure in E. coli. All the active sites mentioned in this article matched with the active sites from *Microcoleus vaginatus.*

**Phosphopyruvate hydratase**

Gene OID: 2505166395 EC: 4.2.1.11

The gene (2505166395) is confirmed to code for the Phosphopyruvate hydratase enzyme (EC 4.2.1.11). A NCBI BLAST [4] (NCBI: Q110V4) search against the swissprot database gives other Phosphopyruvate hydratase enzymes top hits. A PDB BLAST [14] (PDB: 1E9I) search gave an article on the Phosphopyruvate hydratase enzyme from *E. coli*, and eight out of ten of the active sites mentioned in this article matched with the ones in *Microcoleus vaginatus*.

**Glucose-6-phosphate Isomerase**

Gene OID: 2505167736 EC: 5.3.1.9

Gene is confirmed to code for Glucose-6-phosphate Isomerase (GPI). A NCBI BLAST gives as top hits other GPIs, and the e-values and query coverage are both very good [7] (NCBI: B2J5F1). A PDB BLAST gave an article [2] (PDB: 2O2C) about the active site residues for GPI. There were some active site residues that didn’t match, namely T260, F261 and T262. In microcoleus vaginatus, these active site residues were S260, G261 and G262. For site 260, T and S are similar in structure, and the article says that the T is important for its side chain hydroxyl group, which S also has. Also, a multiple sequence alignment was done with FASTA sequences from UniProt, and at site 260 the S residue is completely conserved when compared to other cyanobacteria. For sites 261 and 262, nothing was mentioned in the article, so these sites can be assumed to be less important for the active site conS Figuration. Also, a multiple sequence alignment [23] showed that for a vast majority of cyanobacteria, the GG sequence was conserved.

**Fructose-1,6-bisphosphatase**

Gene OID: 2505169097 and 2505170007 EC: 3.1.3.11

There were two hits in IMG ER for this enzyme number. A BLAST search using gene 2505169097 was done on UniProt, and the results were narrowed down to reviewed results for cyanobacteria. These results were then split into hits for EC 3.1.3.11 only, hits for EC 3.1.3.11 and 3.1.3.37, and hits for EC 3.1.3.37 only. A multiple sequence alignment of each of these subsets was done against the FASTA sequences of genes 2505169097 and 2505170007. The results suggested that gene 2505169097 was most closely correlated to EC 3.1.3.11, while gene 2505170007 was most closely correlated to being bifunctional for EC 3.1.3.11 and 3.1.3.37. As a final check, a phylogenetic tree was made with the FASTA sequences used for the alignments mentioned previously. The resulting tree showed that gene 2505170007 clustered with the bifunctional genes, while gene 2505169097 clustered with the genes for EC 3.1.3.11.

**Pyruvate, water dikinase**

Gene OID 2505165817, 2505167376, and 2505168646 EC: 2.7.9.2

Three genes (2505165817, 2505167376, and 2505168646) were predicted to code for the Pyruvate, water dikinase enzyme (EC: 2.7.9.2). Based on numerical data from COG and Pfam, the enzyme 2.7.9.2 seems to be coded by either 2505165817 or 2505168646, but not 2505167376. PDB BLAST searches [25] (PDB: 2HWG) and NCBI BLASTs [12] (NCBI: Q55905) gave better numerical results for 2505168646 than 2505165817, but both of the active sites matched well for both of these genes from the PDB searches. Based on results from T-Coffee Multiple Sequence Alignment searches and an IMG Genome BLAST search, it is safe to say that genes 2505165817 and 2505168646 are very similar, but we can not be certain that they necessarily bear homology to one another. No clear conclusion can be made regarding gene 2505167376. Further research is needed.

**Phosphoglycerate mutase**

Gene OID 2505170019, 2505169225, 2505165971, and 2505168909 EC 5.4.2.1

**Gene OID 2505170019**

This gene definitely encodes a protein associated with phosphogylcerate mutase behavior, a claim supported by close homologs with other 5.4.2.1 enzymes found in PDB [26] (PDB: 2A6P) and Swissprot databases [6] (NCBI: A4W6B3). However, because the protein has twice as many amino acids as 5.4.2.1 enzymes and multiple alignment via T-coffee leaves a wide region without any matches.

Literature [26] also suggests that distinguishing phosphoglycerate mutase from other homologous enzymes can be problematic, so further research may be necessary completely understanding this gene/process is crucial.

**Gene OID: 2505169225**

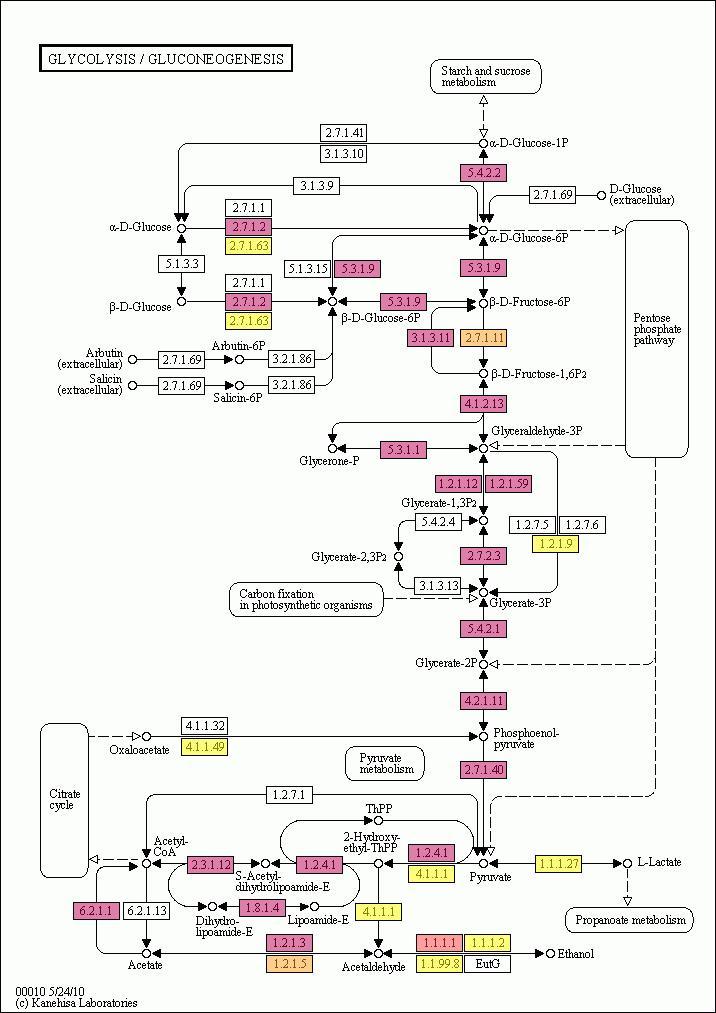
The misannotated gene was found via an IMG Genome BLAST against 2505170019. It was confirmed to code for phosphoglycerate mutase, with reasonably good active site matches on PDB BLAST [28](PDB: 1FZT). An NCBI BLAST with swissprot [29] (NCBI: Q98DM0) revealed other phosphoglycerate mutases in other organisms, all of the biphosphoglycerate-dependent variety.

**Gene OID: 2505165971**

Gene 2505165971 was found to code for a cofactor dependent phosphoglycerate mutase. A NCBI BLAST was done, and the top results were a mixture of alpha-ribazole phosphatases and [probable] phosphoglycerate mutases. A PDB BLAST search [19] (PDB: 1H2E) gave an article about phosphatase PhoE (a member of the cofactor dependent phosphoglycerate superfamily) and the active site residues from this article were compared with the ones in microcoleus vaginatus. Sites 16, 22 and 152 were mismatched. For site 16, N was replaced with S. The article says that N stabilizes the substrate due to its neutral side chain, and S has a neutral side chain as well. Furthermore, N and S are very similar in structure. A multiple sequence alignment was also done on T-Coffee, using FASTA sequences from UniProt, and this alignment showed that the S residue is completely conserved among the compared cyanobacteria sequences. For site 22, according to the article, the presence of a Q established enzyme identity as a PhoE, while the presence of T, S or C established identity as a cofactor dependent phosphoglycerate mutase. A multiple sequence alignment also showed the residue for cyanobacteria was conserved mostly as C, and as S in a few other cases, which is consistent with the article’s findings. Finally, for site 152, the G residue is not mentioned as having a specific function, and a multiple sequence alignment showed that the residue was completely conserved as K for cyanobacteria.

**Gene OID 2505168909**

This gene has been confirmed to encode for 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, the enzyme crucial for reaction 5.4.2.1. This conclusion is supported by very strong evidence for homology with other BPG-independent PGMs in NCBI BLAST and PDB BLAST (PDB:1EQJ). In addition, active sites were a perfect match with the PDB sequence.

**Comparison with Other Genomes**



S Fig. 2:Comparison of genes found in M. vaginatus versus genes found in other cyanobacteria by automatic annotation of glycolysis pathway

Genes from Microcoleus vaginatus were found in most of the other cyanobacteria genomes. There were no new genes found in other cyanobacteria that were not in Microcoleus Vaginatus.

**Functional Studies**

Functional studies could be done on genes 2505166282 and 250516977, which are both predicted to code for EC 4.1.2.13. Two mutants could be grown, each mutant having one gene missing. The two mutants could then be observed to see if they express fructose-bisphosphate aldolase by seeing if increasing the amount of substrate increases the amount of product. It is predicted that each of the mutants would express fructose-bisphosphate aldolase, but if either of them do not, that means one of the genes does not actually code for EC 4.1.2.13.

Functional studies could be done for the genes associated with EC 3.1.3.37 and 3.1.3.11. Gene 2505170007 should be tested to see if it is actually bifunctional. Gene 2505169097 should be repressed and it should be seen if EC 3.1.3.37 is still expressed. Gene 2505170007 should then be repressed separately as well to see if EC 3.1.3.11 is not expressed. If both of these conditions are met, this means gene 2505170007 is bifunctional.

Additionaly, EC 3.1.3.11 takes fructose 1,6-bisphosphate as a substrate, whereas EC 3.1.3.37 takes sedoheptulose 1,7-bisphosphate as its substrate. Therefore, experiments could be done where only one gene is repressed and involving only one of the substrates at a time, to confirm if the genes code for their specified enzymes.

Finally, functional studies could be done on genes 2505165817, 2505167376, and 2505168646, which were all predicted to code for EC 2.7.9.2. Three mutants could be grown, each with 2 of the genes repressed and 1 gene active. It could then be seen if the mutants express EC 2.7.9.2 by seeing if more product is made when more substrate is added. Such a procedure can help to determine if certain genes do or do not code for the enzyme.

**Final Annotation**

An image for the MetaCyc pathway can be found at <http://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=GLYCOLYSIS&detail-level=2&detail-level=1&detail-level=2>

All the enzymes predicted to be in the pathway by MetaCyc were found. Variation I of glycolysis (by MetaCyc’s convention) was used because it says it applies to bacteria. Additionally, there was missing gene for glycolysis III, and glycolysis II is the same as glycolysis I minus the first step of glycolysis I.

**Discussion**

The genes coding for enzymes 5.3.1.9, 3.1.3.11, 2.7.1.11, 5.3.1.1, 1.2.1.12, 2.7.2.3, 5.4.2.1, 4.2.1.11, and 2.7.1.40 were conclusively annotated.

A search for EC 4.1.2.13 within the *Microcoleus vaginatus* genome gave two genes as a result. Both of these genes, 2505166282 and 2505169777, were found to code for the corresponding enzyme. It is not know why there are two genes that code for the same enzyme. Additional studies could be done on M. vaginatus and mutants with only one of the genes to see if having both of these genes confers some sort of evolutionary advantage for the organism.

One enzyme (2505170019) with EC 4.2.1.5 may also need further study if the need arises. It was twice as long as the others and possessed two PG mutase domains; otherwise it aligned nicely in PDB and Swissprot BLAST results. Future study would be needed if EC 4.2.1.5 a more thorough understanding was necessary.

*Microcoleus vaginatus* was found to have a bi-functional gene (2505170007) that coded for both EC 3.1.3.37 and enzyme 3.1.3.11. EC 3.1.3.11 is used in the traditional Calvin Cycle pathway, while EC 3.1.3.37 is used in the alternative pathway. It is uncertain why another gene (2505169097) that only codes for EC 3.1.3.11, since this gene is not needed if gene 2505170007 can already code for EC 3.1.3.11. An experiment could be done where gene 2505169097 is repressed and observations are taken to see if there are any adverse effects on the organism.

Three genes were found to code for enzyme 2.7.9.2 and while two of the genes seem to be very similar to one another, nothing can conclusively said for the remaining gene. The annotation for this enzyme in the document (above) elaborates on this. Further research is needed for this enzyme, specifically for gene 2505167376.

**References**

[1] Anderson,I.J., Dharmarajan,L., Rodriguez,J., Hooper,S., Porat,I.: The complete genome

sequence of Staphylothermus marinus reveals differences in sulfur metabolism among

heterotrophic Crenarchaeota. BMC Genomics 10, 145 (2009)

[2] Arsenieva D, Appavu BL, Mazock GH, Jeffery CJ. Crystal structure of phosphoglucose

isomerase from Trypanosoma brucei complexed with glucose-6-phosphate at 1.6 A

resolution. (2008) Proteins 74:72-80

[3] [Auerbach G](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Auerbach%20G%22%5BAuthor%5D), [Huber R](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Huber%20R%22%5BAuthor%5D), [Grättinger M](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Gr%C3%A4ttinger%20M%22%5BAuthor%5D), [Zaiss K](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Zaiss%20K%22%5BAuthor%5D), [Schurig H](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Schurig%20H%22%5BAuthor%5D), [Jaenicke R](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Jaenicke%20R%22%5BAuthor%5D), [Jacob U](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Jacob%20U%22%5BAuthor%5D). Closed

structure of phosphoglycerate kinase from Thermotoga maritime reveals the catalytic

mechanism and determinants of thermal stability. (1997) Structure 5: 1475-1483

[4] Copeland,A., Lucas,S., Lapidus,A., Barry,K., Detter,J.C., et al. Direct Submission to

Swissprot. Submitted (JUN-2006)

[5] Copeland,A., Lucas,S., Lapidus,A., Barry,K., Detter,J.C.,et al. Swissprot Direct Submission.

Submitted (SEP-2005)

[6] Copeland,A., Lucas,S., Lapidus,A., Barry,K., Glavina del Rio,T., et al. Direct Submission to

Swissprot. Submitted APR-2007

[7] Copeland,A., Lucas,S., Lapidus,A., Glavina del Rio,T., Dalin,E., et al.: Direct Submission to

Swissprot. Submitted APR-2008

[8] Delboni LF, Mande SC, Rentier-Delrue F, Mainfroid V, Turley S, Vellieux FM, Martial JA, Hol

WG.:Crystal structure of recombinant triosephosphate isomerase from Bacillus

stearothermophilus. An analysis of potential thermostability factors in six isomerases with

known three-dimensional structures points to the importance of hydrophobic interactions.

(1995) Protein Sci. 4: 2594-2604

[9] Galkin A, Li Z, Li L, Kulakova L, Pal LR, Dunaway-Mariano D, Herzberg O.: Structural Insights

into the Substrate Binding and Stereoselectivity of Giardia Fructose-1,6- bisphosphate

Aldolase. (2009) Biochemistry 48: 3186-3196

[10] Jedrzejas, M.J., Chander, M., Setlow, P., Krishnasamy, G. “Structure and mechanism of

action of a novel phosphoglycerate mutase form Bacillusstearothermophilus”. (2000)

EMBO J. 19: 1419-1431.

[11]Kaneko,T., Nakamura,Y., Wolk,C.P., Kuritz,T., Sasamoto,S., et al. Complete genomic

sequence of the filamentous nitrogen-fixing cyanobacterium Anabaena sp. strain

PCC 7120. DNA Res. 8 (5), 205-213 2001)

[12] Kaneko,T., Tanaka,A., Sato,S., Kotani,H., Sazuka,T., Miyajima,N.,

Sugiura,M. and Tabata,S. Sequence analysis of the genome of the unicellular

cyanobacterium Synechocystis sp. strain PCC6803. I. Sequence features in the 1 Mb

region from map positions 64% to 92% of the genome. DNA Res. 2 (4), 153-166 (1995)

[13] Klenk,H.P., Clayton,R.A., Tomb,J.F., White,O., Nelson,K.E., et al. The complete genome

sequence of the hyperthermophilic,sulphate-reducing archaeon Archaeoglobus fulgidus.

Nature 390 (6658), 364-370 (1997)

[14] Kühnel K, Luisi BF. Crystal structure of the Escherichia coli RNA degradosome component

enolase. Journal: (2001) J.Mol.Biol. 313: 583

[15] Lorentzen E, Pohl E, Zwart P, Stark A, Russell RB, Knura T, Hensel R, Siebers B. Crystal

Structure of an Archaeal Class I Aldolase and the Evolution of (βα)8 Barrel Proteins. (2003) J.Biol.Chem. 278: 47253

[16] Márquez J, Reinelt S, Koch B, Engelmann R, Hengstenberg W, Scheffzek K.Structure of the

full-length enzyme I of the phosphoenolpyruvate-dependent sugar phosphotransferase

system. Journal: (2006) J.Biol.Chem. 281: 32508-32515

[17] Martin,W., Brinkmann,H., Savonna,C. and Cerff,R.: Evidence for a chimeric nature of

nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate

dehydrogenase genes. Proc. Natl. Acad. Sci. U.S.A. 90 (18), 8692-8696 (1993)

[18] Mattevi A, Valentini G, Rizzi M, Speranza ML, Bolognesi M, Coda A.: Crystal structure of

Escherichia coli pyruvate kinase type I: molecular basis of the allosteric transition. (1995) Structure 3: 729-741

[19] Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res.

2008 Jul 1;36(Web Server issue):W465-9. Epub 2008 Apr 19. (PubMed)

[20] Rigden, D.J., Littlejohn, J.E., Henderson, K., Jedrzejas, M.J. Structures of Phosphate and Trivanadate Complexes of Bacillus Stearothermophilus Phosphatase Phoe: Structural and Functional Analysis in the Cofactor-Dependent Phosphoglycerate Mutase Superfamily (2003) J.Mol.Biol. 325: 411

[21] Schirmer T, Evans PR.: Structural basis of the allosteric behaviour of phosphofructokinase.

(1990) Nature 343: 140-145

[22] Suzuki K, Ito S, Shimizu-Ibuka A, Sakai H.Crystal structure of pyruvate kinase from

Geobacillus stearothermophilus. (2008) J.Biochem. 144: 305-312

[23] T-Coffee: A novel method for fast and accurate multiple sequence alignment. (2000) J Mol

Biol. 302(1):205-17.

[24] The UniProt Consortium. Ongoing and future developments at the Universal Protein

Resource Nucleic Acids. Res. 39: D214-D219 (2011).

[25] Teplyakov, A., Lim, K., Zhu, P.P., Kapadia, G., Chen, C.C., et al. Structure of phosphorylated

enzyme I, the phosphoenolpyruvate:sugar phosphotransferase system sugar

translocation signal protein. Journal: (2006) Proc.Natl.Acad.Sci.Usa 103: 16218-16223

[26] Watkins HA, Baker EN. Structural and functional analysis of Rv3214 from Mycobacterium

tuberculosis, a protein with conflicting functional annotations, leads to its characterization

as a phosphatase. (2006) J.Bacteriol. 188: 3589-3599

[27] Yun M, Park CG, Kim JY, Park HW.: Structural analysis of glyceraldehyde 3-phosphate

dehydrogenase from Escherichia coli: direct evidence of substrate binding and

cofactor-induced conformational changes. (2000) Biochemistry 39: 10702-10710

[28] [Uhrínová S](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Uhr%C3%ADnov%C3%A1%20S%22%5BAuthor%5D), [Uhrín D](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Uhr%C3%ADn%20D%22%5BAuthor%5D), [Nairn J](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Nairn%20J%22%5BAuthor%5D), [Price NC](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Price%20NC%22%5BAuthor%5D), [Fothergill-Gilmore LA](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Fothergill-Gilmore%20LA%22%5BAuthor%5D), [Barlow PN](http://www.ncbi.nlm.nih.gov/pubmed?term=%22Barlow%20PN%22%5BAuthor%5D).:Solution

structure and dynamics of an open beta-sheet, glycolytic enzyme, monomeric 23.7 kDa phosphoglycerate mutase from Schizosaccharomyces pombe.” (2001) J.Mol.Biol. 306: 275-290

[29] Kaneko,T., Nakamura,Y., Sato,S., Asamizu,E., Kato,T., Sasamoto,S.: Complete

genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti. DNA Res. 7 (6), 331-338 (2000)

# Hydrogenase

No hydrogenase encoding gene is predicted (via homology) within the *Microcoleus vaginatus* PCC 9802 genome. It is possible that some other nitrogen-fixing bacteria in the desert crust is responsible for the aerobic photoproduction of hydrogen or that Microcoleus possesses a very unique hydrogenase. Further experimentation may be necessary to verify this.

Hydrogen photoproduction by soil crust communities may be achieved by a number of means. In many bacteria nitrogenase is the hydrogen-evolving enzyme in cyanobacteria as well as in phototrophic bacteria [[1](http://onlinelibrary.wiley.com/doi/10.1111/j.1399-3054.1982.tb00584.x/pdf)], which may be triggered under anaerobic conditions. In addition, photosystem I has also been implicated in hydrogen production of cyanobacteria [[2](http://onlinelibrary.wiley.com/doi/10.1111/j.1751-1097.1985.tb01609.x/pdf)]. Known nitrogen-fixing bacteria capable of aerobic hydrogen photoproduction include *Anabaena variabilis* and *Nostoc sp.*[3, [4](http://www.biomedcentral.com/1756-0500/4/186)].

Non-nitrogen fixing cyanobacteria like *Anacystis nidulans* (*Synechoccus elongates* PCC 7942) have also been found to possess aerobic and anaerobic hydrogenase. Genes encoding hydrogenase nickel incorporation protein hypA (P94160) and HoxH NAD-reducing hydrogenase HoxS (P94159) from *Synechococcus* PCC 6301 were BLAST searched against the *Microcoleus* genome. One weak homolog was found for hypA (gene OID 2505169900: E-value 6e-01, 30% identity), yet when a BLAST search of that gene was done against the Swissprot database, stronger homology was found with a verified leukotoxin (P16462, E-value 6e-20, 39% identity) [[5](http://www.ncbi.nlm.nih.gov/pubmed/2670940)].

A couple weak homologs of the HoxS gene were found via BLAST in the *Microcoleus* genome. One gene, OID 2505168283, (MV BLAST result: E-value 1e-02, 22% identity) was predicted by IMG to code for phosphoribosylformylglycinamidine cyclo-ligase and a BLAST search in the Swissprot database supported this prediction; strong homology with a verified phosphoribosylformylglycinamidine cyclo-ligase was found (B7KJK8, E-value 3e-169, 72% identity). Another gene, OID 2505167087, that IMG identified as uncharacterized (BLAST result: E-value 8e-02, 24% identity) also had a non-hydrogenase homolog in Swissprot (verified phosphoglycerate mutase: P36623, E-value 1e-23, 34% identity) [[6](http://www.ncbi.nlm.nih.gov/pubmed/8110200)].

IMG/ER also predicted one gene (2505169669) to code for “Fe-S-cluster-containing hydrogenase components 2” due to homology with COG1142: “Fe-S-cluster-containing hydrogenase components 2” (E value 7.0e-17, 22.9% identity). A BLAST search against PDB and Swissprot databases revealed only weak homologs like “Electron transport complex protein rnfB” (C5BDE6, E-value 0.004, 40% identity). This protein most likely has iron-binding behavior, but it is hard to say if it does encode hydrogenase without further experimentation.

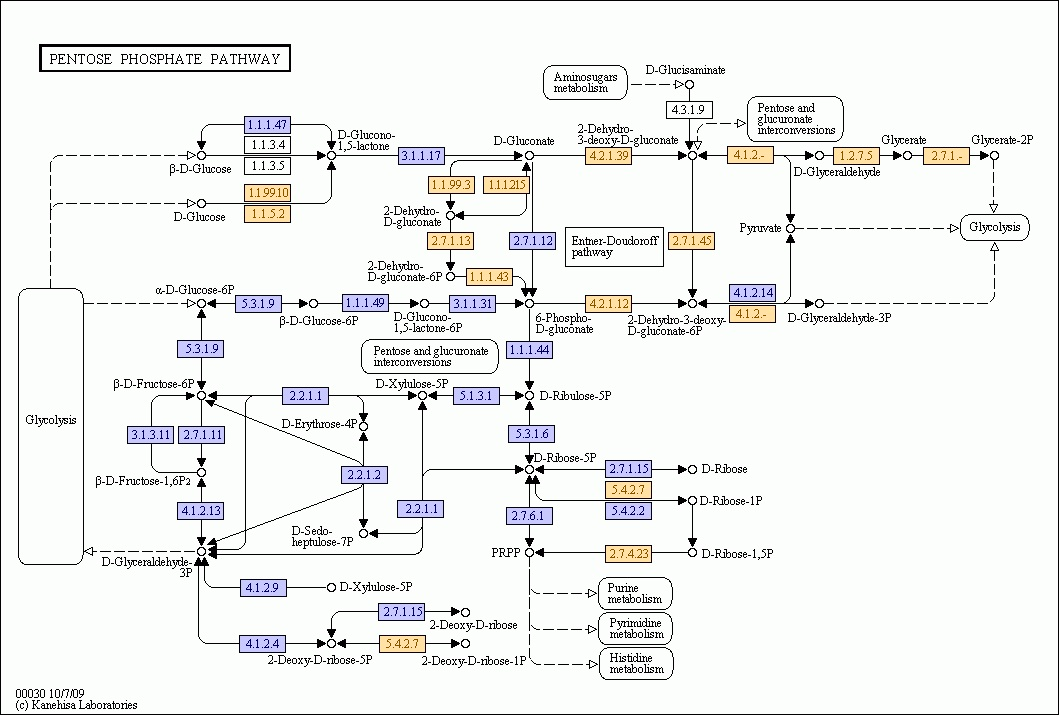
**References:**

1. Kerfin, W. and P. Boger (1982) *Physiol. Plant* **54**, 93–98.
2. Randt C, Senger H (1985) Participation of the two photosystems in light dependent hydrogen evolution in *Scenedesmus obliquus*. Photochem Photobiol 42:553–557
3. S. Tsygankov, L. T. Serebryakova, D. A. Sveshnikov, K. K. Rao, I. N. Gogotov, D. O. Hall. Hydrogen photoproduction by three different nitrogenases in whole cells of Anabaena variabilis and the dependence on pH, International Journal of Hydrogen Energy, Volume 22, Issue 9, September 1997, Pages 859-867, ISSN 0360-3199, DOI: 10.1016/S0360-3199(96)00242-X.<http://www.sciencedirect.com/science/article/pii/S036031999600242X>)
4. M. Holmqvist, P. Lindberg, Å. Agervald, K. Stensjö, P. Lindblad. Transcript analysis of the extended *hyp*-operon in the cyanobacteria *Nostoc* sp. strain PCC 7120 and *Nostoc punctiforme* ATCC 29133. *BMC Research Notes* 2011, 4:186doi:10.1186/1756-0500-4-186
5. E. Lallys, E. Golub, I. Kieba, N. Taichman, J. Rosenbloom, J. Rosenbloom, C. Gibson, D. Demuth. Analysis of the Actinobacillus actinomycetemcomitans Leukotoxin Gene. Journal of Biological Chemistry. (1989) Vol . 264, No. 2 6 , Issue of September 15., pp. 15451-15456,1989.
6. Nairn J, Price NC, Fothergill-Gilmore LA, Walker GE, Fothergill JE, et al. The amino acid sequence of the small monomeric phosphoglycerate mutase from the fission yeast*Schizosaccharomyces pombe*. *Biochem J.* 1994;297 (Pt 3):603–608.

# Pentose Phosphate Pathway

We annotated the Pentose Phosphate Cycle for *Microcoleus vaginatus*. We consulted the MetaCyc Pentose Phosphate pathway to see which enzymes were specific to the pathway, and we tried to find genes in *Microcoleus vaginatus* that coded for those enzymes. The pathway is important since it allows for the oxidation of glucose in order to make NADPH and other carbohydrates to be used for biosynthesis. This pathway can be used as an alternative to glycolysis. The Pentose Phosphate pathway consists of two phases. In the irreversible oxidative phase, glucose-6-phosphate is converted into ribulose-5-phosphate through oxidative decarboxylation. This phases results in the NADPH generation. The next phase is the reversible and non-oxidative phase. It allows for phosphorylated sugars to be interconverted to generate xylulose-5-phosphate, ribulose-5-phosphate, and ribose-5-phosphate.

**Automated annotation [KEGG + Metacyc]**



S Fig. 1: Map of Pentose Phosphate Cycle and predicted genes (blue) and predicted missing genes (orange) in M. vaginatus

Metacyc link: <http://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PENTOSE-P-PWY&detail-level=2>

**Glucose-6-phosphate 1-dehydrogenase**

Gene OID: 2505169095 EC: 1.1.1.49

Gene is confirmed to code for glucose-6-phosphate 1-dehydrogenase, with almost no ambiguity. NCBI BLAST [6](NCBI: P73411) reveals near-perfect matches with other G6PD enzymes, including from the cyanobacterium Synechosystis, and PDB BLAST [4](PDB: 1E77) reveals a closely related G6PD in *L. Mesenteroides* with a perfect active site match.

**6-Phosphogluconolactonase**

Gene OID: 2505170610 EC: 3.1.1.31

Gene was confirmed to code for 6-phosphogluconolactonase. A NCBI BLAST [1](NCBI: P46016) search gave as top hits other 6-phosphogluconolactonases with a high degree of homology. A PDB BLAST [5](PDB: 2J0E) search gave an article on the 6-phosphogluconolactonase structure from *Trypanosoma brucei*. A check of the active site residues showed that all active site residues matched between the ones described in the article and the ones in *Microcoleus vaginatus*.

**6-Phosphogluconate Dehydrogenase, Decarboxylating**

Gene OID: 2505169298 EC: 1.1.1.44

Gene was confirmed to code for 6-phosphogluconate dehydrogenase. A NCBI BLAST [7](NCBI: P52208) search was done and the top hits were other 6-phosphogluconate dehydrogenases. The e-values for these hits were very low and the query coverage was high. Additionally, a PDB BLAST [2](PDB: 2W8Z) search was done and an article on the 6-phosphogluconate dehydrogenase structure in *Geobaccilus stearothermophilus* was found. The active sites in this article were compared to the active site residues in *Microcoleus vaginatus*, and all of them matched.

**Ribulose-phosphate 3-epimerase**

Gene OID: 2505170722 EC: 5.1.3.1

Gene was confirmed to code for ribulose-phosphate 3-epimerase. A NCBI BLAST [8](NCBI: P74061) search gave as top results other ribulose-phosphate 3-epimerases, and the matches were very good according to the e values and query coverage. A PDB BLAST [14](PDB: 1TQJ) search gave an article on the ribulose-phosphate 3-epimerase structure in *Synechocystis sp.*, a cyanobacteria. The alignment showed that the match was very good, and additionally all the checked active sites matched up with each other.

**Ribose 5-Phosphate Isomerase**

GeneOID: 2505165655 EC: 5.3.1.6

This enzyme and the gene coding for it were already found in the Calvin Cycle annotation. This entry is a copy of the description from the Calvin Cycle Annotation Report.

BLAST searches within IMG/ER and UniProt both gave ribose 5-phosphate isomerases as the top hits, as did NCBI BLAST [9](NCBI: Q111U2). returned an extremely good match with transketolase. A PDB BLAST [13] (PDB: 1LK5) search gave as a 1st result the article "A hyperthermostable D-ribose-5-phosphate isomerase from Pyrococcus horikoshii characterization and three-dimensional structure. (2002) Structure 10: 877-886".

The enzyme number matched and the organism used was *Pyrococcus horikoshii*.

The active sites were checked, and they all matched except for sites 100, 106 and 168.

In the organism, the sites were R100, M106 and D168. In *Micrcoleus vaginatus*, the sites were G100, R106 and D168. In the article, there was no mention of the functions of sites 106 or 168. It was said that site 100 was not completely conserved. Also, a multiple sequence alignment was showed in the article. This alignment showed that the aforementioned sites were not completely conserved, and the amino acids found in *Micrcoleus vaginatus* were found in some of the other organisms used for the alignment. To double-check, a multiple sequence alignment was also done against other similar genes from cyanobacteria. Site 100 was almost completely conserved along cyanobacteria as being a G, except for one cyanobacterium that was a T. Site 106 had many R's among other cyanobacteria, as well as Q and a single A. Site 168 had mostly A's among cyanobacteria, a couple D's, and one T. The few outliers could be explained as not actually being the correct enzyme, because some cyanobacteria listed multiple genes for enzyme number 5.3.1.6, which could be a mistake due to automatic annotation.

**Bacterial Transketolase**

Gene OID: 2505170664 EC: 2.2.1.1

This enzyme and the gene coding for it were already found from the Calvin Cycle annotation. This entry is a copy of the one from the Calvin Cycle Annotation Report. PDB BLAST revealed this enzyme to be transketolase, with an e-value of 0 and a perfect active site match [](PDB: 1L6W). NCBI BLAST with Swissprot came to the same conclusion with far greater certainty (E-value: ~0.0) [3](NCBI: Q118F4).

This gene most likely encodes for the 2.2.1.1 reaction enzyme due to homology and 100% active site matching with yeast transkelotase, which has EC 2.2.1.1. It also bears strong homology to Swissprot verified transkelotase.

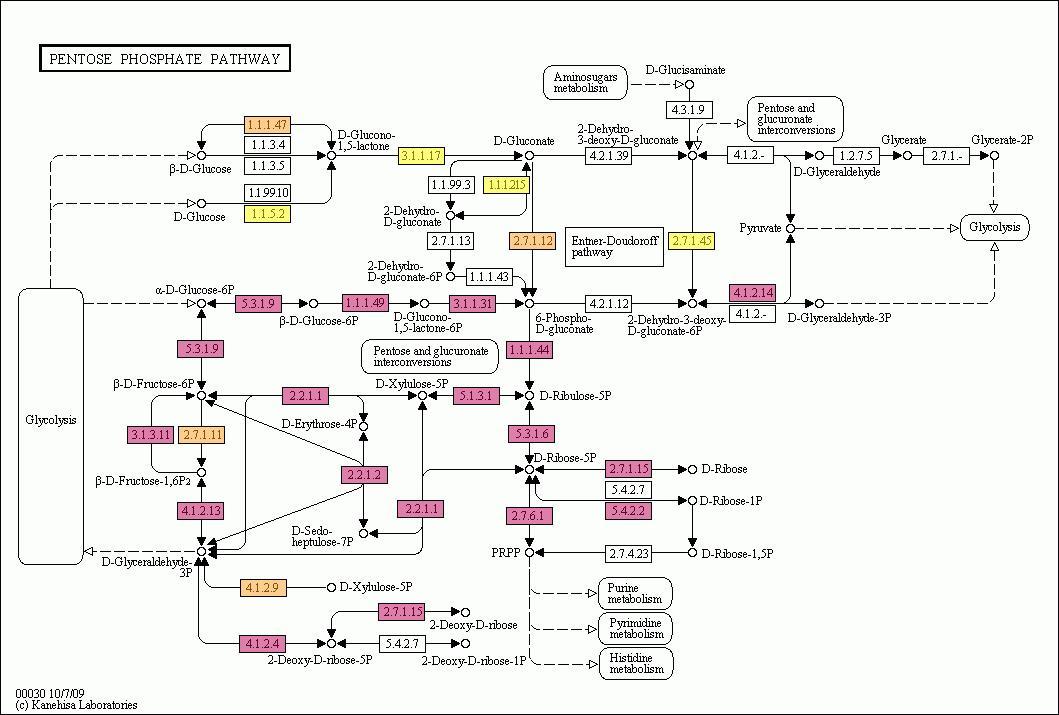
**Transaldolase**

Gene OID: 2505170982 and 2505169096 EC: 2.2.1.2

There were two different genes that showed up as hits for this enzyme. Gene 2505170982 was predicted to be a fructose-6-phosphate aldolase while gene 2505169096 was predicted to be a transaldolase.

Gene 2505170982 was predicted to be a fructose-6-phosphate aldolase on the basis of an article from a PDB BLAST search [12](PDB: 1L6W). This article had a list of active site residues that were conserved for both aldolases and transaldolases. All these active sites matched with the ones in microcoleus vaginatus. The article also had a list of active site residues that were different between aldolases and transaldolases. Out of these active sites, 5/8 sites matched the sites for an aldolase. No sites matched the sites for a transaldolase. The other 3 sites were mismatches. A mismatch at site 107, L to A, could be explained by the fact that these two amino acids are similar in structure. At site 57, F was replaced by Y, and they are similar except for a hydroxyl group on the aromatic ring. Finally, at site 129, A was replaced by I, and they are similar in that they are both carbon based side chains, only of different lengths. From this data, it was concluded that the gene was incorrectly annotated and should actually be a fructose-6-phosphate aldolase. Also, a NCBI BLAST search gave as top hits both transaldolases and fructose-6-phosphate aldolases [10](NCBI: B4U8P1), so the results from the PDB article were the deciding factor.

Gene 2505169096 was predicted to be a transaldolase on the basis of NCBI and PDB BLASTs along with an InterProScan search. InterProScan predicted the gene to code for a transaldolase. NCBI BLAST gave other transaldolases as top hits, and PDB BLAST brought up articles about transaldolases as top hits [11](PDB: 1I2N).

**Comparison with Other Genomes**



S Fig. 2: Comparison of genes found in M. vaginatus versus genes found in other cyanobacteria by automatic annotation of the Pentose Phosphate Cycle

Genes from *Micrcoleus vaginatus* were found in other cyanobacteria genomes. There were no new genes found in other cyanobacteria that were not in *Micrcoleus vaginatus*.

**Functional Studies**

Functional studies could be done on two of the genes, 2505170982 and 2505169096, which were both predicted to code for EC 2.2.1.2. 2505170982 was predicted to actually code for fructose-6-phosphate aldolase, while 2505169096 was predicted to code for transaldolase. Two mutants of M. vaginatus could be observed, each mutant having one of the genes missing. It should be seen whether increasing the amount of substrate for the mutants results in an increasing in products; this will help to determine if the genes code for fructose-6-phosphate aldolase or transaldolase.

**Final Annotation**

Metacyc link: <http://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PENTOSE-P-PWY&detail-level=2>

All the enzymes predicted by MetaCyc were found to be coded for in the genome of *Micrcoleus vaginatus*.

**Discussion**

All of the enzymes except for EC: 2.2.1.2 were found to have genes that coded for them unambiguously. EC: 2.2.1.2, transaldolase, had some ambiguity in its coding gene. Two genes, 2505170982 and 2505169096, were found as hits for this enzyme. It was concluded that the first gene coded for fructose-6-phosphate aldolase while the second coded for transaldolase, althought there was some uncertainty in this prediction. If time permits, this prediction could be tested in the lab. Gene 2505169096 could be suppressed and if *Micrcoleus vaginatus* could not go through the pentose phosphate cycle, this would mean that the gene indeed codes for transaldolase.

EC 5.1.3.1 and 5.3.1.6 look to be redundant on first glance, since it seems that the two pathways ultimately do the same thing: they convert D-ribulose-5-phosphate to D-sedoheptulose-7-phosphate, just through different intermediates. Further examination of the MetaCyc pathway however shows that EC 2.2.1.1 requires both D-ribulose-5-phosphate and D-sedoheptulose-7-phosphate as substrates; it does not take just one or the other substrate, but takes both at the same time.

**References**

[1] Bauer,C.C., Scappino,L. and Haselkorn,R. Direct Submission. Submitted (??-SEP-1994)

[2] Cameron, S., Martini, V.P., Iulek, J., Hunter, W.N. *Geobacillus stearothermophilus*6-phosphogluconate dehydrogenase complexed with 6-phosphogluconate. Journal: (2009) Acta Crystallogr.,Sect.F 65: 450-454

[3] Copeland,A., Lucas,S., Lapidus,A., Barry,K., Detter,J.C., et al. Direct Submission. Submitted (??-JUN-2006)

[4] Cosgrove, M.S., Gover, S., Naylor, C.E., Vandeputte-Rutten, L., Adams, M.J., Levy, H.R. An examination of the role of asp-177 in the His-Asp catalytic dyad of *Leuconostoc mesenteroides*glucose 6-phosphate dehydrogenase: X-ray structure and pH dependence of kinetic parameters of the D177N mutant enzyme. Journal: (2000) Biochemistry 39: 15002

[5] Delarue, M., Duclert-Savatier, N., Miclet, E., Haouz, A., et al. Three dimensional structure and implications for the catalytic mechanism of 6-phosphogluconolactonase from *Trypanosoma brucei*. Journal: (2007) J.Mol.Biol. 366: 868

[6] Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., et al. Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3 (3), 109-136 (1996)

[7] Kaneko,T., Tanaka,A., Sato,S., Kotani,H., Sazuka,T., Miyajima,N., Sugiura,M. and Tabata,S. Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92% of the genome. DNA Res. 2 (4), 153-166 (1995)

[8] Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., et al. Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3 (3), 109-136 (1996)

[9] Kaneko,T., Nakamura,Y., Wolk,C.P., Kuritz,T., Sasamoto,S., et al. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium Anabaena sp. strain PCC 7120. DNA Res. 8 (5), 205-213 (2001)

[10] Reysenbach,A.L., Hamamura,N., Podar,M., Griffiths,E., Ferreira,S., et al. Complete and draft genome sequences of six members of the Aquificales. J. Bacteriol. 191 (6), 1992-1993 (2009)

[11] Schorken, U., Thorell, S., Schurmann, M., Jia, J., Sprenger, G.A., Schneider, G.Identification of catalytically important residues in the active site of Escherichia coli transaldolase. (2001) Eur.J.Biochem. 268: 2408-2415

[12] Thorell, S., Schurmann, M., Sprenger, G.A., Schneider, G.Crystal structure of decameric fructose-6-phosphate aldolase from *Escherichia coli* reveals inter-subunit helix swapping as a structural basis for assembly differences in the transaldolase family. (2002) J.Mol.Biol. 319: 161-171

[13] Wikner, C., Nilsson, U., Meshalkina, L., Udekwu, C., Lindqvist, Y., Schneider, G. Identification of catalytically important residues in yeast transkelotase. : (1997) Biochemistry 36: 15643-15649

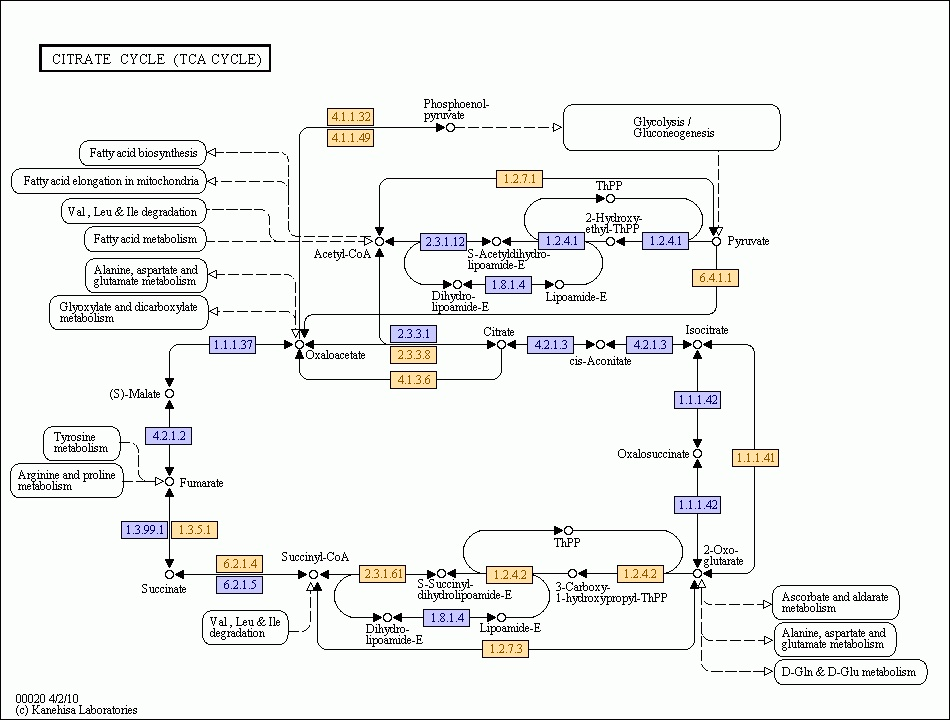
[14] Wise, E.L., Akana, J., Gerlt, J.A., Rayment, I. Structure of D-ribulose 5-phosphate 3-epimerase from Synechocystis to 1.6 A resolution. Journal: (2004) Acta Crystallogr.,Sect.D 60: 1687-1690

# TCA Cycle

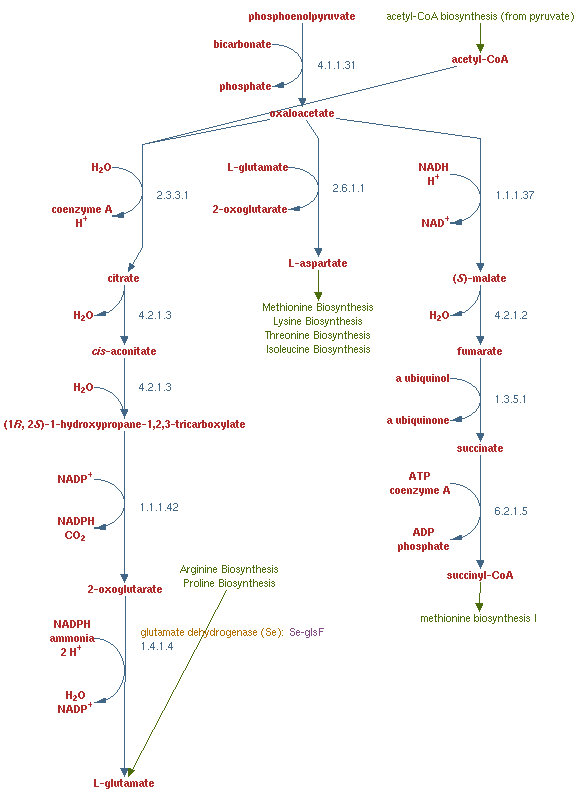
We annotated the tricarboxylic acid (TCA) cycle pathway for *M. vaginatus* PCC 9802. The TCA cycle, also known as the citrate cycle or the Krebs cycle, is an important part of cellular respiration. However, cyanobacteria such as *Microcoleus vaginatus* do not have a complete TCA cycle since they lack α-ketoglutarate dehydrogenase or EC:1.2.4.2 [25]. This was the old understanding of cyanobacteria, but another reference showed that M. vaginatus can complete its cycle and convert 2-oxoglutarate to succinate through an alternative pathway called the GABA shunt [9]. However, one of the enzymes needed for this alternate pathway was not found, so it is unclear if M. vaginatus can undergo the GABA shunt..

According to MetaCyc, when an organism has a complete TCA cycle, it generates energy and replenishes electron carriers. For cyanobacteria however, the cycle does not close and only serves to produce amino acids or succinyl-CoA, which can subsequently be synthesized into other amino acids.

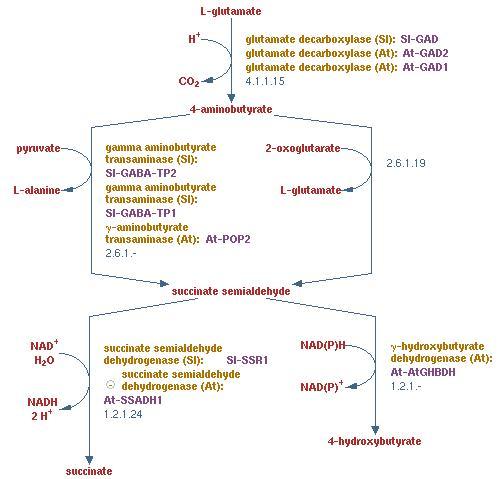
**Automated Annotation [from KEGG and MetaCyc]**



S Fig. 1: The TCA cycle in cyanobacteria is incomplete. In particular, enzyme 1.2.4.2 (α-ketoglutarate dehydrogenase) does not exist in *Microcoleus vaginatus*. Also, a separate pathway involving aspartate transaminase (2.6.1.1) is not shown in the diagram.



S Fig. 2: The incomplete TCA cycle in cyanobacteria



S Fig. 3: The GABA Shunt for M. vaginatus, which follows the left pathway and replaces EC 1.2.1.24 with 1.2.1.16.

**2-methylcitrate synthase/citrate synthase II**

Gene OID: 2505168309 EC: 2.3.3.1

The gene is confirmed to code for enzyme 2.3.3.1. NCBI BLAST [14] (NCBI: Q59977) reveals other citrate synthetases as excellent matches. Interestingly, all the top matches from PDB [23] (PDB: 1AJ8) reveal that *Microcoleus*’ copy of this gene is very closely related to thermophilic bacteria and archaea such as *Thermus*, *Sulfolobus*, and *Pyrococcus*. All the active site residues were matches or near-matches (structurally similar amino acid replacements). This suggests that this gene is probably adapted for the extreme ground temperatures that *Microcoleus* endures.

**Succinyl CoA Synthetase (2 subunits)**

Gene OID’s: 2505166031 and 2505166030 EC: 6.2.1.5

This protein has two subunits that are coded for by two genes with IMG/ER gene OID’s 2505166031 and 2505166030. These two genes are confirmed to code for the subunits. NCBI BLAST (alpha subunit: [7] (NCBI:P45102)) (beta subunit: [24] (NCBI: Q0VPF7)) ) reveals other succinyl CoA synthetases as matches, and the top PDB BLAST hits (alpha subunit: [10] (PDB: 2NU7)) (beta subunit: [11] (PDB: 1CQI)) reveals a very similar enzyme in *E. coli*.The active sites match in the alpha subunit (the beta subunit seems to be slightly different in structure, though the evidence indicates that the *E. coli* protein and this one are homologs).

**Aconitase**

Gene OID: 2505168762 EC: 4.2.1.3

Gene is confirmed to code for enzyme 4.2.1.3. A NCBI BLAST search [13] (NCBI: P74582) against the Swissprot database gives other aconitases as top hits. A PDB BLAST search [32] (PDB: 1L5J) gave an article on E. coli aconitase structure, and all the active sites mentioned in this article matched with the ones in *Microcoleus vaginatus*.

**Isocitrate Dehydrogenase (NADP-dependent)**

Gene OID: 2505168569 EC: 1.1.1.42

Gene is confirmed to code for enzyme 1.1.1.42. A NCBI BLAST search [12] (NCBI: P50214) against the Swissprot database gives other isocitrate dehydrogenases as top hits. A PDB BLAST search [19] (PDB: 1AI2) gave an article isocitrate dehydrogenase structure, and all but one of the active sites mentioned in this article matched with the ones in *Microcoleus vaginatus*. The article says the active site at 345 should be a tyrosine, but for *Microcoleus vaginatus* it is a histidine. A multiple sequence alignment was performed on T-Coffee [28], which showed the histidine residue to be completely conserved at site 345 for other cyanobacteria. The FASTA sequences used for comparison were obtained from UniProt.

**Glutamate dehydrogenase (NADP Specific)**

Gene OID: 2505169473 EC: 1.4.1.4

Gene is confirmed to code for enzyme 1.4.1.4. A NCBI BLAST search [2] (NCBI: P54386) against the swissprot database gives other glutamate dehydrogenases (NADP specific) as top hits. A PDB BLAST search [21] (PDB: 1EUZ) gave an article on glutamate dehydrogenase structure, and all the active sites mentioned in the article matched with the ones in microcoleus vaginatus.

**Malate Dehydrogenase**

Gene OID: 2505168756 EC: 1.1.1.37

Gene is confirmed to code for enzyme 1.1.1.37. A NCBI BLAST search [12] (NCBI: Q8YP78) gave other malate dehydrogenases as top hits. A PDB BLAST search [4] (PDB: 1GUZ) gave as a top hit an article on the structure of malate dehydrogenases. From this article, the active site residues were compared, and all but one of them matched. The mismatch at site 10 involved the replacing of an Asn with a Lys. According to the article, the side chain nitrogen group Asn 10 interacts through hydrogen bonding with the two phosphate groups in one of the substrates, NAD. In microcoleus vaginatus, Lys 10 could also feasibly interact through hydrogen bonding, because Lys also has a side chain nitrogen group.

As another check, a multiple sequence alignment was done on T-Coffee [28], using FASTA sequences [29] of malate dehydrogenases in other cyanobacteria . The results showed that at site 10, most cyanobacteria either had Asn or Arg as the amino acid. Some did have Lys at site 10. According to T-Coffee, the match between Asn, Arg and Lys shows up as good, and the hypotheses from earlier is confirmed because these three amino acids all have side chain nitrogen groups. It seems plausible that either Asn 10, Arg 10 or Lys 10 would be acceptable active sites for this enzyme.

**Fumerase/Aspartate Ammonia-Lyase**

Gene OID: 2505169957 EC: 4.2.1.2/4.3.1.1

The gene most likely codes for the predicted enzyme 4.2.1.2. A PDB BLAST [31] (PDB: 2FUS) suggests that this enzyme is fumarase. Similarly, NCBI BLAST [26] (NCBI: P26899) suggests that this enzyme is fumerate hydratase. However, because it bears strong sequence homology to Aspartate ammonia-lyase (EC 4.3.1.1) there are the possibilities that this gene is bifunctional or encodes for 4.3.1.1, not 4.2.1.2. Despite this, because Asparatate ammonia-lyase and fumerase belong to the same enzyme family (Aspartate ammonia-lyase), are suspected to make use of the common reaction mechanism for catalysis, and studies characterizing active sites are sparse, meaning this decision may require further research.

**Phosphoenolpyruvate carboxylase**

Gene OID: 2505170379 EC: 4.1.1.31

The gene (2505170379) was confirmed to code for EC 4.1.1.31. A NCBI BLAST [3] (NCBI: Q3MB82) search against the swissprot database gives other Phosphoenolpyruvate carboxylase as top hits. A PDB BLAST search [17] (PDB: 1JQN) gave an article on the Phosphoenolpyruvate carboxylase structure, and all the active sites but one mentioned in the article matched with the ones in *Microcoleus vaginatus*.

**Aspartate transaminase**

Gene OID’s: 2505169488 and 2505171077 EC: 2.6.1.1

Two genes (2505169488 and 2505171077) were predicted to code for this enzyme. NCBI BLAST searches for each of the genes [14] (NCBI:Q55128) [5] (NCBI:O67781) against the swissprot database gave other Aspartate transaminase enzymes as top hits. PDB BLAST searches for each of the genes [20] (PDB: 1BJW) [30] (PDB: 1B5P) gave articles on Aspartate transanimase, and for both of the articles gave active site residues; all these residues matched for both of the genes. An IMG BLAST Search also provided reliable numerical data that both the genes were homologs. The genes were found to be homologs to each other, with 25.14 percent identity and an e value of 1e-25.

**Succinate dehydrogenase or Fumarate reductase (Multiple Subunits)**

Gene OID’s: 2505168142 and 2505166800 EC: 1.3.5.1/1.3.99.1 (ambiguous)

These genes most likely code for subunits of the Succinate dehydrogenase enzyme. In subunit A (OID: 2505168142), a PDB BLAST [34] (PDB:1NEK) returns succinate dehydrogenase subunit as the top hit, as does NCBI BLAST [33] (NCBI:Q9UTJ7). The same applies to subunit B (OID: 2505166800) where PDB BLAST [22] (PDB:2WP9) and NCBI BLAST [18] (NCBI:Q8ZQU2) both strongly suggested that this gene coded for another subunit of the enzyme. Although the Metacyc pathway for the FCA cycle variation IV requires an enzyme for reaction 1.3.5.1, 1.3.99.1 performs the same task and is predicted in other variations of the pathway (the only difference is 1.3.99.1 can act with any electron acceptor). Furthermore, while the succinate dehydrogenase for 1.3.5.1 requires four subunits, only two possible subunits are found in M. Vaginatus. This leaves the distinction between a gene encoding 1.3.5.1 and 1.3.99.1 ambiguous and perhaps needs further clarification in the lab. This should be kept in mind for subsequent annotations for the 3 other subunits of succinate dehydrogenase.

A third subunit, Gene OID 2505168273, heterodisulfide reductase, was predicted to possess functions 1.3.99.1 and 1.3.98.1 but was not homologous with other 1.3.99.1 enzymes. It is possible that this gene’s function had a false prediction, but it is also possible that the enzyme in M. vaginatus needs to be studied further.

**Glutamate decarboxylase**

Gene OID: 2505167018 EC: 4.1.1.15

Gene was confirmed to code for glutamate decarboxylase. A NCBI BLAST search [27] (NCBI: P48321) against the Swissprot database gave other glutamate decarboxylases as top hits. A PDB BLAST search [1] (PDB: 2OKK) gave as a top hit an article about glutamate decarboxylase. Finally, a search was done on UniProt [29] for reviewed genes in bacteria coding for EC 4.1.1.15. Their FASTA sequences were taken and then a multiple sequence alignment was done on T-Coffee [28], which gave a good amount of matching between the M. vaginatus gene and the genes from other organisms.

**Succinate-semialdehyde dehydrogenase (NAD(P)(+))**

Gene OID: 2505170539 EC: 1.2.1.16

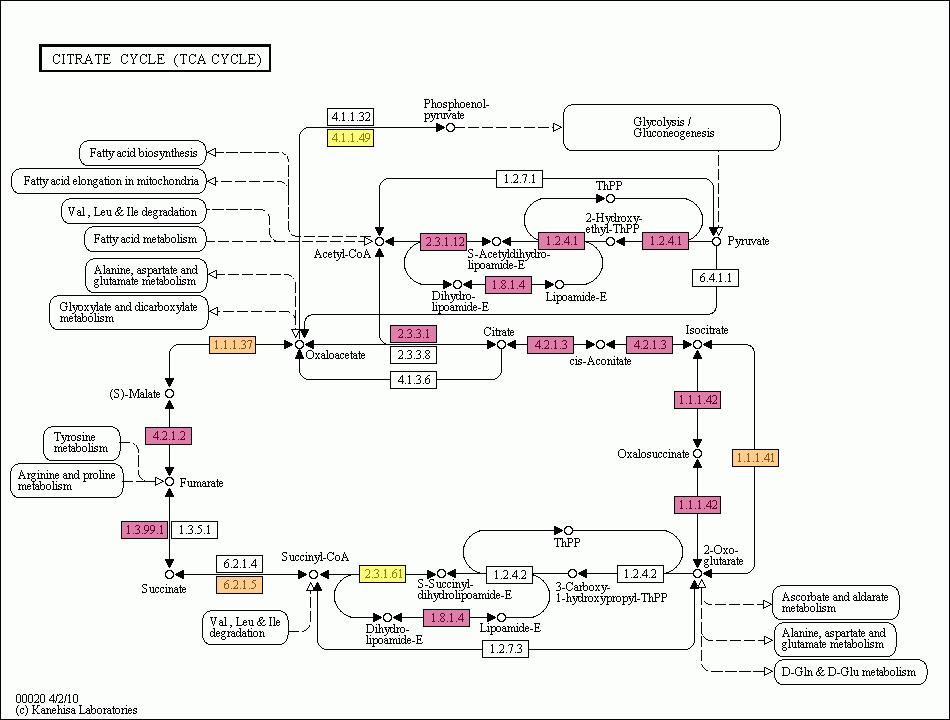
Gene is confirmed to code for succinate-semialdehyde dehydrogenase. A NCBI BLAST search [16] (NCBI: Q73TP5) gave as top hits other succinate-semialdehyde dehydrogenases with a high degree of confidence. A PDB BLAST search [15] (PDB: 3JZ4) gave an article on succinate-semialdehyde dehydrogenase structure. 12/17 of the active sites in the article matched with those in M. vaginatus. The mismatches were at sites 153, 239, 243, 282 and 338. The mismatches were changes from T, Q, Q, R, K to M, S, N, L, D. In the article, these sites were not mentioned in great detail, nor were any extremely important functions associated with them. As an additional check, a multiple sequence alignment was done on T-Coffee [28]. The gene from M. vaginatus was compared with genes from other cyanobacteria found in UniProt [29] to code for the same EC number. This alignment showed very good matching between the different genes. Furthermore, it showed the M residue at site 153 to be competely conserved, the S residue at site 239 to be mostly conserved, the L residue at site 282 to be mostly conserved, and the D residue at site 338 is conserved among some of the other cyanobacteria genes. Only site 243 did not match well, with 243 having mostly A for the other cyanobacteria.

**4-aminobutyrate transaminase**

Gene OID: 2505170763 EC: 2.6.1.19

The gene that was found as a hit for this EC number actually showed up as a hit for EC 2.6.1.11 because IMG/ER did not predict this gene to exist in M. vaginatus. There is ambiguity because it seems that the gene could code for either EC 2.6.1.11 or 2.6.1.19. A PDB BLAST search gave two articles as results. One was a hit for EC 2.6.1.11 [24] (PDB: 2BP0) while the other was a hit for EC 2.6.1.19 [17] (PDB: 1SF2). Each article mentioned active site residues, and the gene from M. vaginatus showed good matches for the active site residues from both articles. To further clarify, a phylogenetic tree [23] was made to see if the gene clustered more with EC 2.6.1.11 or 2.6.1.19. The FASTA sequences used for comparison were obtained from UniProt [29]. The 2.6.1.11 sequences were obtained by searching for EC 2.6.1.11 and narrowing the results down to reviewed results for cyanobacteria. The 2.6.1.19 sequences were obtained by searching for EC 2.6.1.19 and narrowing the results down to reviewed results for bacteria. The phylogenetic tree result showed that all the 2.6.1.11 and 2.6.1.19 clustered separately into two distinct clusters, and the tree showed that the M. vaginatus gene clustered with the 2.6.1.11 genes. It seems that the gene codes for EC 2.6.1.11, but more research should be done to see if perhaps the gene is bifunctional or to see if the gene is misannotated.

**Comparison with Other Genomes**



S Fig. 4: Comparison of genes found in M. vaginatus versus genes found in other cyanobacteria by automatic annotation of the TCA Cycle

Genes from Microcoleus vaginatus were found in most of the other cyanobacteria genomes. There were no new genes found in other cyanobacteria that were not in Microcoleus Vaginatus.

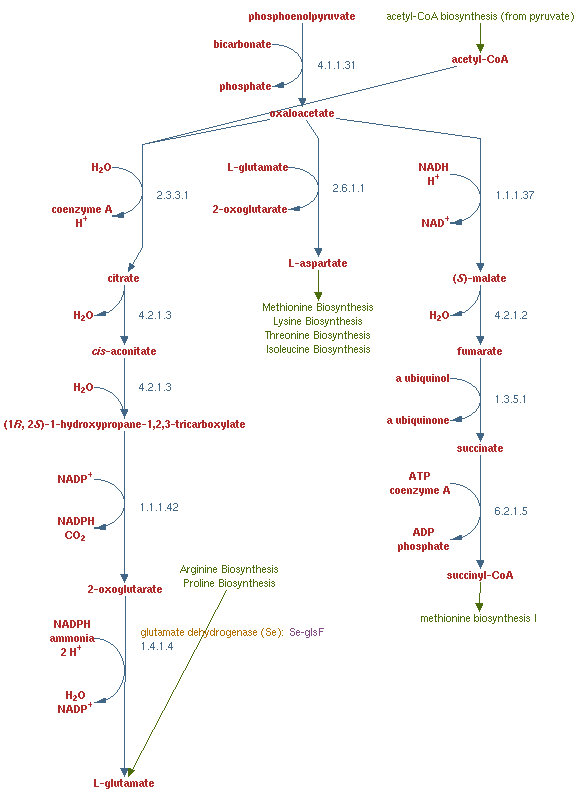
**Functional Studies**

Functional studies can be done on gene 2505169957. This gene was predicted to code for EC 4.2.1.2, but it could possibly code for EC 4.3.1.1 instead or as well. Both enzymes would catalyze their substrates to the same product, fumarate, but they act on different substrates. EC 4.2.1.2 acts on malate while EC 4.3.1.1 acts on aspartate. According to the MetaCyc map, both of these substrates come from oxaloacetate by different pathways. Therefore, experiments could be run with only one of the pathways working, so only aspartate or malate is produced. It could then be observed whether or not fumarate is still produced. If fumarate is produced from either aspartate or malate, the gene is bifunctional for both enzymes. If fumarate is produced only from malate, then the gene codes for EC 4.2.1.2, and if fumarate is produced only from aspartate, then the gene codes for EC 4.3.1.1.

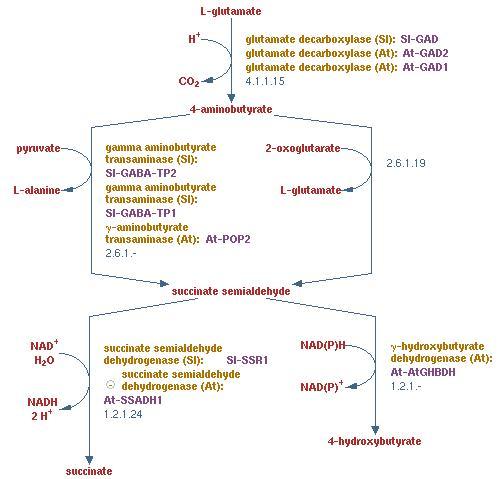
Functional studies can also be done on genes 2505168142 and 2505166800. These genes are predicted to code for either EC 1.3.5.1 or 1.3.99.1, which are both succinate dehydrogenases, but with different substrates. EC 1.3.5.1 uses ubiquinone as the electron acceptor while EC 1.3.99.1 any electron acceptor, not just a specific one. To test the genes and see if they are EC 1.3.5.1 or 1.3.99.1, multiple experiments should be done: one with ubiquinone as the electron acceptor and other ones with other electron acceptors. If the reaction can only be catalyzed by the enzyme in the presence of ubiquinone, the gene codes for EC 1.3.5.1. If the reaction can be catalyze by the enzyme in the presence of any of the electron acceptors, the gene codes for EC 1.3.99.1.

Finally, functional studies should be done on gene 2505170763 to see if it codes for EC 2.6.1.19, or to see if any other gene codes for EC 2.6.1.19 or 2.6.1.-. Either one of these enzymes is needed to complete the GABA shunt. An experiment should first be done to see if an enzyme exists that helps catalyse the reaction of 4-aminobutyrate to succinate semialdehyde. It should be seen if increasing levels of 4-aminobutyrate increases the amount of succinate semialdehyde present. If it does, it confirms that the enzyme exists and there is a gene that codes for it. Further studies would then need to be done to find the gene that codes for the enzyme.

**Final Annotation**



S Fig. 5: The incomplete TCA cycle for cyanobacteria



S Fig. 6: The GABA Shunt for M. Vaginatus, which hypothetically goes along the right pathway and replaces EC 1.2.1.24 with 1.2.1.16

All the genes coding for enzymes in the incomplete TCA cycle pathway were found and annotated. The genes coding for enzymes in the GABA shunt were found and annotated as well.

**Discussion**

The genes coding for enzymes 2.3.3.1, 4.2.1.3, 1.1.1.42, 1.4.1.4, 4.1.1.31, and 1.1.1.37 were successfully annotated with a high degree of certainty.

Succinyl CoA synthetase (6.2.1.5) has two subunits. The alpha subunit, which contains the main active site, very closely resembles that of E. Coli. However, the beta subunit, which binds ADP, exhibits a considerable degree of difference in its active site area. Though the beta subunit is almost certainly homologous to the corresponding beta subunit in E. Coli, this difference is rather interesting.

The enzyme for reaction 4.2.1.2 bears strong homology for an enzyme in reaction 4.3.1.1. Although this may just be due the common ancestry of the proteins (in the same family) and common reactions, further research may be necessary to determine exactly which function this enzyme holds--if it is bifunctional or one of the two.

IMG/ER predicted two genes (2505169488 and 2505171077) to code for the Aspartate transaminase enzyme (EC 2.6.1.1). Since all numerical results (COG, Pfam, PDB) proved to be reliable for both genes, an IMG Genome BLAST Search was carried out to determine the homology between the two genes. The numerical data (e-value and percent identity) from this search was reliable, leading one to believe that two genes may bear homology. However, further research must be carried out to confidently conclude that the two genes are homologs of one another.

Reaction 1.3.5.1 was interesting because no cyanobacteria possessed the enzymes necessary. 1.3.5.1 involves Succinate dehydrogenase using ubiquinone as an electron acceptor, and is commonly undertaken in mitochondria. However, function 1.3.99.1, which uses a component of the Sdh enzyme in 1.3.5.1, was predicted in M. vaginatus genes. The difference is that 1.3.99.1 can use other electron acceptors besides ubiquinone (which is prominent in mitochondria) and 1.3.99.1 possess 2 subunits instead of 4. It would be interesting to see if EC 1.3.5.1 somehow involved in function from 1.3.99.1, such that it gained substrate specificity to ubiquinone. It would also be interesting to see if there is some evolutionary benefit to using only ubiquinone as an electron acceptor instead of any electron acceptor.

Interestingly, many of the genes closely resemble thermophilic bacteria and archaea such as Thermus, Thermococcus, Sulfolobus, and Pyrococcus. This indicates that many of the genes in this metabolic pathway are probably adapted to extreme temperatures, which is consistent with the extreme ground temperatures that are characteristic of the Sonoran desert.

Finally, although M. vaginatus does not have a traditional TCA cycle, since it has no 2-oxoglutarate dehydrogenase to catalyze the reaction of 2-oxoglutarate to succinyl-CoA and eventually succinate, there is an alternative pathway called the GABA shunt [9]. This effectively allows M. vaginatus to convert 2-oxoglutarate into succinate. This pathway operates by converting 2-oxoglutarte into glutamate, converting glutamte into 4-aminobutyrate, convering 4-aminobutyrate into succinate semialdehyde, and finally converting succinate semialdehyde into succinate, but it is not certain that the pathway exists in M. vaginatus. An experiment in the lab is needed to verify or refute the presence of the GABA shunt.

**References**

[1] Buckle, A.M., Fenalti, G., Law, R.H.P., Whisstock, J.C.: GABA production by glutamic

acid decarboxylase is regulated by a dynamic catalytic loop. (2007) Nat.Struct.Mol.Biol.

14: 280-286.

[2] Chavez,S., Reyes,J.C., Chauvat,F., Florencio,F.J. and Candau,P.: The NADP-glutamate

dehydrogenase of the cyanobacterium Synechocystis 6803: cloning, transcriptional

analysis and disruption of the gdhA gene. Plant Mol. Biol. 28 (1), 173-188 (1995)

[3]Copeland,A., Lucas,S., Lapidus,A., Barry,K., Detter,J.C.,et al. Swissprot Direct Submission.

Submitted (SEP-2005)

[4] Dalhus B, Saarinen M, Sauer UH, Eklund P, Johansson K, Karlsson A, Ramaswamy S,

Bjørk A, Synstad B, Naterstad K, Sirevåg R, Eklund H.: Structural Basis for Thermophilic

Protein Stability: Structures of Thermophilic and Mesophilic Malate

Dehydrogenases (2002)

[5] Deckert,G., Warren,P.V., Gaasterland,T., Young,W.G., Lenox,A.L., et al.: The complete

genome of the hyperthermophilic bacterium Aquifex aeolicus. Nature 392 (6674),

353-358 (1998)

[6] Ewart,G.D., Reed,K.C. and Smith,G.D. Soluble hydrogenase of Anabaena cylindrica. Cloning

and sequencing of a potential gene encoding the tritium exchange subunit. . J. Biochem.

187 (1), 215-223 (1990).

[7] Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., et al.: Whole-genome random

sequencing and assembly of Haemophilus influenzae Rd. Science 269 (5223), 496-512

(1995)

[8] Han, G.W., Schwarzenbacher, R., Page, R., Jaroszewski, L., Abdubek, P., Ambing, E.

Crystal structure of an alanine-glyoxylate aminotransferase from Anabaena sp. at 1.70 A

resolution reveals a noncovalently linked PLP cofactor. Journal: (2005) Proteins 58:

971-975.

[9] Henning Knoop, Yvonne Zilliges, Wolfgang Lockau, and Ralf Steuer: The Metabolic Network

of Synechocystis sp. PCC 6803: Systemic Properties of Autotrophic Growth.

[10] Hidber E, Brownie ER, Hayakawa K, Fraser ME.: Participation of Cys 123 alpha of

Escherichia coli Succinyl-CoA Synthetase in Catalysis. (2007) ACTA

CRYSTALLOGR.,SECT.D 63: 876-884

[11] Joyce MA, Fraser ME, James MN, Bridger WA, Wolodko WT.: ADP-binding site of

Escherichia coli succinyl-CoA synthetase revealed by x-ray crystallography. (2000)

Biochemistry 39: 17-25

[12] Kaneko,T., Nakamura,Y., Wolk,C.P., Kuritz,T., Sasamoto,S., et al. Complete genomic

sequence of the filamentous nitrogen-fixing cyanobacterium Anabaena sp. strain PCC 7120. DNA Res. 8 (5), 205-213 2001)

[13] Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., et al.: Sequence

analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3 (3), 109-136 (1996)

[14] Kaneko,T., Tanaka,A., Sato,S., Kotani,H., Sazuka,T., Miyajima,N.,

Sugiura,M. and Tabata,S. Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92% of the genome. DNA Res. 2 (4), 153-166 (1995)

[15] Langendorf, C.G., Key, T.L.G., Fenalti, G., et al.: The X-ray crystal structure of

Escherichia coli succinic semialdehyde dehydrogenase; structural insights into

NADP+/enzyme interactions. (2010) Plos One 5: e9280-e9280.

[16] Li,L., Bannantine,J.P., Zhang,Q., Amonsin,A., May,B.J., Alt,D., Banerji,N., Kanjilal,S. and

Kapur,V.: The complete genome sequence of Mycobacterium avium subspecies  
 paratuberculosis. Proc. Natl. Acad. Sci. U.S.A. 102 (35), 12344-12349 (2005).

[17] Liu, W., Peterson, P.E., Carter, R.J.: Crystal structures of unbound and

aminooxyacetate-bound Escherichia coli gamma-aminobutyrate aminotransferase.

(2004) Biochemistry 43: 10896-10905.

[18] Matsumura H, Xie Y, Shirakata S, Inoue T, Yoshinaga T, Ueno Y, Izui K, Kai Y.: Crystal

structures of C4 form maize and quaternary complex of E. coli phosphoenolpyruvate

carboxylases. Journal: (2002) Structure 10: 1721-1730

[19] McClelland,M., Sanderson,K.E., Spieth,J., Clifton,S.W., et al.: Complete genome sequence

of Salmonella enterica serovar Typhimurium LT2. Nature 413 (6858), 852-856 (2001)

[20] Mesecar AD, Stoddard BL, Koshland DE Jr.: Orbital steering in the catalytic power of

enzymes: small structural changes with large catalytic consequences. (1997)

Science 277: 202-206

[21] Nakai T, Okada K, Akutsu S, Miyahara I, Kawaguchi S, Kato R, Kuramitsu S, Hirotsu K.:

Structure of Thermus thermophilus HB8 aspartate aminotransferaseand its

complex with maleate. (1999) Biochemistry 38: 2413-2424

[22] Nakasako M, Fujisawa T, Adachi S, Kudo T, Higuchi S. Large-scale domain movements

and hydration structure changes in the active-site cleft of unligated glutamate

dehydrogenase from Thermococcus profundus studied by cryogenic X-ray crystal structure analysis and small-angle X-ray scattering. (2001) Biochemistry 40: 3069-3079

[23] Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008

Jul 1;36(Web Server issue):W465-9. Epub 2008 Apr 19. (PubMed)

[24] Rajaram, V., Ratna Prasuna, P., Savithri, H.S., Murthy, M.R.N.: Structure of biosynthetic

N-acetylornithine aminotransferase from Salmonella typhimurium: Studies on substrate

specificity and inhibitor binding. (2007) Proteins 70**:** 429-441.

[25] Ruprecht J, Iwata S, Rothery RA, Weiner JH, Maklashina E, Cecchini G. Perturbation of the

quinone binding site of complex II alters the electronic properties of the

proximal[3Fe-4S] iron-sulfur cluster. (2011) J.Biol.Chem. 286: 12756-12765

[26] Russell, R.J., Ferguson, J.M., Hough, D.W., Danson, M.J., Taylor, G.L.: The crystal

structure of citrate synthase from the hyperthermophilic archaeon pyrococcus furiosus at

1.9 A resolution. (1997) Biochemistry 36: 9983-9994.

[27] Schneiker,S., Martins dos Santos,V.A., Bartels,D., Bekel,T., et al.: Genome sequence of the

ubiquitous hydrocarbon-degrading marine bacterium Alcanivorax borkumensis. Nat.

Biotechnol. 24 (8), 997-1004 (2006)

[28] Stanier, R.Y. and G. Cohen-Bazire. “Phototrophic Prokaryotes: the Cyanobacteria.” Annu.

Rev. Microbiol. 1977.31:256-257.

[29] Sun,D.X. and Setlow,P.: Cloning, nucleotide sequence, and expression of the Bacillus

subtilis ans operon, which codes for L-asparaginase and L-aspartase. J. Bacteriol.

173 (12), 3831-3845 (1991)

[30] Suzuki,R., Asami,N., Amann,E. and Wagatsuma,M.: Sequences of two porcine glutamic

acid decarboxylases (65- and 67-kDa GAD). Gene 152 (2), 257-260 (1995).

[31] T-Coffee: A novel method for fast and accurate multiple sequence alignment. (2000) J Mol

Biol. 302(1):205-17.

[32] The UniProt Consortium. Ongoing and future developments at the Universal Protein

Resource Nucleic Acids. Res. 39: D214-D219 (2011).

[33] Ura H, Nakai T, Kawaguchi SI, Miyahara I, Hirotsu K, Kuramitsu S.: Substrate recognition

mechanism of thermophilic dual-substrateenzyme. (2001) J.BIOCHEM.(TOKYO) 130:

89-98

[34] Weaver T, Lees M, Banaszak L.: Mutations of Fumerase that distinguish between the

active site and a nearby dicarboxylic acid binding site. (1997) Protein Sci. 6:

834-842

[35] Williams CH, Stillman TJ, Barynin VV, Sedelnikova SE, Tang Y, Green J, Guest JR,

Artymiuk PJ.: E. coli aconitase B structure reveals a HEAT-like domain with implications

for protein-protein recognition. (2002) Nat.Struct.Biol. 9: 447-452

[36] Wood,V., Gwilliam,R., Rajandream,M.A., Lyne,M., Lyne,R., Ussery,D., Barrell,B.G.

and Nurse,P: The genome sequence of Schizosaccharomyces pombe. Nature 415

(6874), 871-880 (2002)

[37] Yankovskaya V, Horsefield R, Törnroth S, Luna-Chavez C, Miyoshi H, Léger C, Byrne B,

Cecchini G, Iwata S. Architecture of succinate dehydrogenase and reactive oxygen

species generation. (2003) Science 299: 700-704

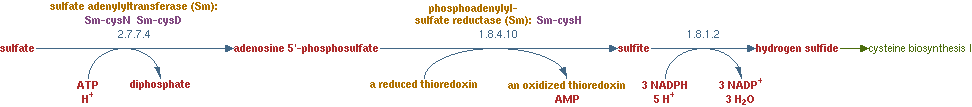
# Sulfur and Sulfate Reduction

The ability to reduce sulfur is not found in all organisms, but rather is limited to those organisms that have the sulfur reductase enzyme (EC 1.97.1.3). This enzyme catalyzes the reduction of sulfur to hydrogen sulfide. The reaction also involves diatomic hydrogen, which donates an electron in order to reduce sulfur. Organisms which have this sulfur reductase enzyme are known as sulfur-reducing organisms.

A search was done on IMG/ER to see if there were any genes from *Microcoleus vaginatus* predicted to code for EC 1.97.1.3. No genes were predicted, so a search was done on UniProt for any genes that were assigned to code for an enzyme with EC 1.97.1.3 function. Three such genes were found, and their FASTA sequences were BLASTed against the *Microcoleus vaginatus* genome from IMG/ER. This BLAST search gave no hits for sulfur reductase in the *Microcoleus vaginatus* genome. Finally, a search was done by enzyme name for any sulfur reductases predicted be IMG/ER to be in the *Microcoleus vaginatus* genome. No hits were given as a result. Therefore, it is predicted that *Microcoleus vaginatus* does not have the sulfur reductase enzyme, and thus cannot reduce sulfur and so is not a sulfur-reducing organism.

Although *Microcoleus vaginatus* was not found to have the ability to reduce sulfur, it was found to have the ability to reduce sulfate. This organism uses sulfate as a sulfur source, reducing sulfate to hydrogen sulfide through the sulfate reduction pathway. The initial step in sulfate reduction is its activation to adenosine 5'-phosphosulfate (APS), which is catalyzed by the enzyme sulfate adenylyltransferase (ATP sulfurylase). Two types of enzymes are known to catalyze the direct reduction of APS to sulfite - a glutathione -dependent enzyme (EC 1.8.4.9), and a thioredoxin -dependent enzyme (EC 1.8.4.10). Sulfite is reduced to sulfide by the action of sulfite reductase. Again, two types of enzymes are known - a ferredoxin-dependent enzyme (EC 1.8.7.1) and an NADPH-dependent enzyme (EC 1.8.1.2). Most bacteria utilze NADPH as the electron donor. *Microcoleus vaginatus* utilizes the thioredoxin dependent enzyme (EC 1.8.4.10) and the NADPH dependent enzyme (EC 1.8.1.2) in this pathway.

**Automated Annotation**



S Figure 1: MetaCyc illustration of sulfate reduction pathway.

**Sulfate adenylyltransferase**

**Gene OID: 2505165890 EC: 2.7.7.4**

This gene is predicted to code for sulfate adenylyltransferase. A NCBI BLAST search gave only sulfate adenylyltransferases and bifunctional suflate adenylyltransferases as hits. The top published hit had an e-value of 0 and 75% identity [[5]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-5). Additionally, a PDB BLAST search gave as a top hit an article on sulfate adenylyltransferase crystal structure [[6]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-6) with an e-value of 1.38863E-72 and 43% identity. Out of the active sites mentioned in the article, all 7 of them matched. Also, a reverse BLAST search gave this gene as the top hit. From this consideration, the perfect active site matching and the strong NCBI and PDB BLAST hits, this gene is strongly predicted to code for sulfate adenylyltransferase.

**Sulfite reductase (NADPH dependent)**

**Gene OID: 2505170520 EC: 1.8.1.2**

This gene was initially predicted by IMG/ER to code for sulfite reductase ferredoxin (EC 1.8.7.1), but BLAST searches and active site matching suggest that this was a misannotation. A NCBI BLAST search gave sulfite reductase ferredoxin as the top two hits [[4]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-4) with e-values of 0 and around 60% identity, but the rest of the hits were to sulfite reductase (NADPH dependent). The best hit had an e-value of 1E-146 and 40% identity (NCBI: C0Z9X2). The hit to sulfite reductase hemoprotein was more reliable, because it was backed up by a paper about sulfite reductase hemoprotein. On the other hand, the hit to sulfite reductase ferredoxin was backed up by a paper about the general genome of a cyanobacteria, and the assignment of functions to its genes.

Additionally, a PDB BLAST search gave hits only to articles on sulfite reductase hemoprotein. The best hit had an e-value of 2.20209E-88 and 36% identity [[2]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-2). Active sites were mentioned in this article, and out of 23 active sites mentioned, 19 of them matched. The four mismatches were N223Q, S258G, S436A and F437M. Three of the mismatches, N223Q, S258G and F437M, were not mentioned in the article and most likely are not essential for catalytic function. Also, N223 and Q223 are similar in structure. Residue S436 was mentioned for its ability to form a strained backbone and hydrogen bond with its carbonyl peptide. A436 is very similar to S436, only differing by a hydroxyl group, and A436 can also hydrogen bond with its carbonyl peptide. On top of all this, a reverse BLAST search gave this gene as the top hit, so it is predicted to code for NADPH dependent sulfite reductase.

**Adenylyl-sulfate reductase**

**Gene OID: 2505168049 EC: 1.8.4.10**

This gene was initially predicted by IMG/ER to code for phosphoadenosine phosphosulfate reductase (EC 1.8.4.8). A NCBI BLAST search gave as the top result a hit to a phosphoadenosine phosphosulfate reductase with an e-value of 7E-106 and 73% identity [[3]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-3). Most of the NCBI BLAST hits were to phosphoadenosine phosphosulfate reductases; however, some hits were to adenylyl sulfate reductase. The best such hit had an e-value of 7E-106 and 63% identity [[3]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-3). At this point, it was unclear which enzyme the gene coded for.

However, a PDB BLAST search gave an article on adenylyl-sulfate reductase as the top result [[1]](http://metaboliteatlas.org/converter/converter_nodb.php#ref-1) with an e-value of 8.30084E-94 and 66% identity. Out of the 31 active sites mentioned in the article, 29 of them matched. The only 2 mismatches were I145V and R164K. The I145V mismatch is not a major one, because I145 and V145 are very similar in structure. The same can be said of the R164K mismatch, as R164 and K164 are also both similar in structure, both having long side chains with nitrogens on them. Also, a reverse BLAST search was done and it gave this gene as the top hit, so it is strongly predicted to code for adenylyl-sulfate reductase.

**Comparison with Other Genomes**

A comparison was done on IMG/ER to see if other cyanobacteria were found to have the enzymes needed for the sulfate reduction pathway. This search predicted all but one of the cyanobacteria in the database to have a gene to code for sulfate adenylyltransferase (EC: 2.7.7.4). However, most of the cyanobacteria in the database were not predicted to have a gene coding for sulfite reductase (NADPH) (EC: 1.8.1.2), and none were predicted to have a gene coding for adenylyl-sulfate reductase (thioredoxin) (EC: 1.8.4.10). This at first seems problematic, since it seems to suggest that no other cyanobacteria can undergo this sulfate reduction pathway, but this was actually the same result as for *Microcoleus vaginatus*. The organism was predicted to only have sulfate adenylyltransferase (EC: 2.7.7.4), and the other enzymes were found, misannotated as other enzymes. Sulfite reductase (NADPH) (EC: 1.8.1.2) was misannotated as sulfite reductase (ferredoxin) (EC: 1.8.7.1) and adenyly-sulfate reductase (thioredoxin) (EC: 1.8.4.10) was misannotated as phosphoadenylyl-sulfate reductase (thioredoxin) (EC: 1.8.4.8).

A search was done to see if cyanobacteria had the EC 1.8.48 or EC 1.8.7.1 enzymes. This search showed that all but one of the cyanobacteria in the IMG/ER database had both of these enzymes. It is therefore hypothesized that these gene predictions are also misannotations, such that most cyanobacteria do have the enzymes necessary for the sulfate reduction pathway outlined previously.

One thing that should be noted is that one of the misannotated enzymes is very similar to the actual enzyme. Sulfite reductase (NADPH) and sulfite reductase (ferredoxin) only differ in the electron carrier but otherwise catalyze the same reaction. Therefore, even if the misannotation theory is wrong, most cyanobacteria still have two of the enzymes needed for this pathway. Although it would still potentially be missing the last enzyme, it would not make much sense for the organism to have two of the enzymes for the pathway yet not complete the pathway; as such, a search could be done and could turn up misannotated or hidden results for the last enzyme.

**Discussion**

*Microcoleus vaginatus* is predicted to be able to undergo sulfate reduction, and most cyanobacteria are predicted to be able to undergo this pathway as well. *Microcoleus vaginatus* is therefore not unique in its ability to reduce sulfate. It makes sense for *Microcoleus vaginatus* to be able to reduce sulfate itself, since this would make it less dependent on other organisms. If it was not able to reduce sulfate itself, it would be dependent on other organisms which could reduce sulfate.

**References**

[1] Chartron J, Carroll KS, Shiau C, Gao H, Leary JA, Bertozzi CR, Stout CD. "Substrate

recognition, protein dynamics, and iron-sulfur cluster in Pseudomonas aeruginosa

adenosine 5'-phosphosulfate reductase." J Mol Biol. Nov 24, 2006; 364(2):152-69.

[2] Crane BR, Siegel LM, Getzoff ED. "Structures of the siroheme- and Fe4S4-containing active

center of sulfite reductase in different states of oxidation: heme activation via

reduction-gated exogenous ligand exchange." Biochemistry. Oct 7, 1997;

36(40):12101-19.

[3] Gutierrez-Marcos JF, Roberts MA, Campbell EI, Wray JL. "Three members of a novel small

gene-family from Arabidopsis thaliana able to complement functionally an Escherichia

coli mutant defective in PAPS reductase activity encode proteins with a thioredoxin-like

domain and "APS reductase" activity." Proc Natl Acad Sci U S A. Nov 12, 1996;

93(23):13377-82.

[4] Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M,

Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo

K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M,

Tabata S. "Sequence analysis of the genome of the unicellular cyanobacterium

Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and

assignment of potential protein-coding regions." DNA Res. Jun 30, 1996; 3(3):109-36.

[5] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[6] Taguchi Y, Sugishima M, Fukuyama K. "Crystal structure of a novel zinc-binding ATP

sulfurylase from Thermus thermophilus HB8." Biochemistry. Apr 13, 2004;

43(14):4111-8.

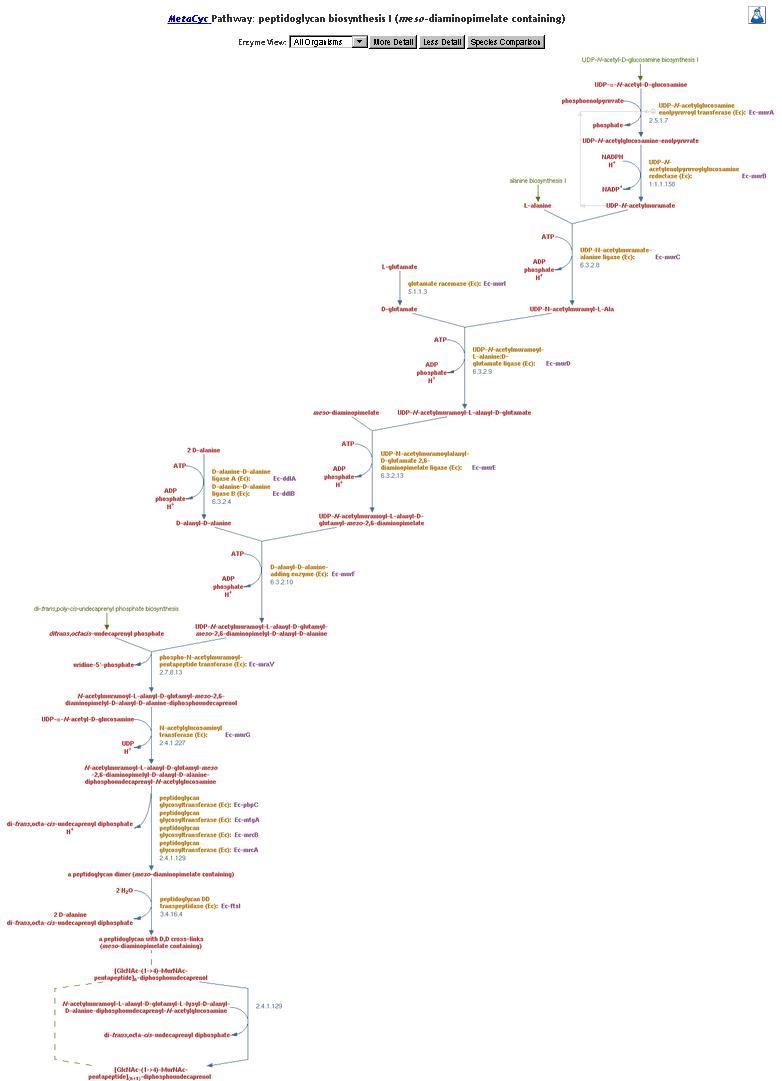
# Peptidoglycan Biosynthesis

The bacterial bag-shaped envelope (sacculus) can resist turgor pressures of 2-5 atmpspheres in Gram-negative bacteria, and up to 20 atm in Gram-positive bacteria. This remarkable tensile strength, combined with elasticity, is achieved by the structure of the cell wall's main component, peptidoglycan (also known as "basal structure", "mucopeptide", "glycopeptide", "glycosaminopeptide" and "murein"). Peptidoglycan is found on the outside of the cytoplasmic membrane of almost all eubacteria, and is unique to these organisms.

Peptidoglycan is best described as a fisherman net. The mesh of the net is made of two segments of parallel, rather inextensible glycan threads, held together by two small elastic peptide crosslinks allowing the net to expand or shrink. The glycan moiety of the peptidoglycan is remarkably uniform among all bacteria, and is made up of alternating β-1,4-linked *N*-acetylglucosamine and *N*-acetyl muramate residues, with an average chain length (in different organisms) of 10 to 65 disaccharide units (although it is about 5000 in *Bacillus subtilis*).

In other words, peptidoglycan biosynthesis is basically the synthesis of cell walls. This pathway describes how the cell walls are made in cyanobacteria, and presumably in *Microcoleus vaginatus*. All of the genes predicted by MetaCyc to be necessary for the pathway were present, which makes sense because *Microcoleus vaginatus* is observed to be an organism that has cell walls.

**Automated Annotation**



S Fig. 1: MetaCyc illustration of peptidoglycan biosynthesis (full size picture from http://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=PEPTIDOGLYCANSYN-PWY&detail-level=2)

**UDP-N-acetylmuramoylalanine--D-glutamate ligase**

Gene OID: 2505167555 EC: 6.3.2.9

Gene is predicted to code for EC 6.3.2.9. A NCBI BLAST search [6] gave as top results other UDP-N-acetylmuramoylalanine--D-glutamate ligases, with very low e-values (around 1E-171) and high percent identity (around 65%). A PDB BLAST search [9] gave an article on UDP-N-acetylmuramoylalanine--D-glutamate ligase with an e-value of 1.2479E-49 and 31% identity. This article mentioned 8 conserved active site residues, and they all matched with the *Microcoleus vaginatus* gene. A multiple sequence alignment was also done on T-Coffee with reviewed cyanobacteria UDP FASTA sequences from UniProt. The alignment showed very strong matching between gene 2505167555 and the ones from other cyanobacteria.

**D-alanine--D-alanine ligase**

Gene OID: 2505168929 EC: 6.3.2.4

Gene is predicted to code for EC 6.3.2.4. A NCBI BLAST search [8] gave as top results other D-alanine--D-alanine ligases with very low e-values (1E-155) and high percent identity (around 70%). A PDB BLAST search [10] gave an article on D-alanine--D-alanine ligase, with an e-value of 2.52151E-79 and 45% identity. The active site residues were checked, and 19 out of 22 of them matched. The mismatches were Q214 to A214, L94 to V94 and M310 to F310. According to the article, Q214 helped to contribute a hydrogen bond with its main chain oxygen, but any amino acid can contribute a hydrogen bond with its main chain oxygen. L94 is only mentioned to be special because its side chain is nonpolar. V94 also has a nonpolar side chain. M310 was also only mentioned to be special because of its nonpolar side chain, and F310 also has a nonpolar side chain.

As a second check, a multiple sequence alignment was done on T-Coffee, using reviewed cyanobacteria FASTA sequences from UniProt. This showed that at site 214, other cyanobacteria have Q214 and A214, so it is justifiable that Q214 is replaced by A214 for gene 2505168929. At site 94, showed some cyanobacteria have L94 and V94, and a lot of other amino acids at the residue. Finally, site 310 showed no other cyanobacteria with F310, but as mentioned in the article, it seems that any amino acid with a nonpolar side chain should preserve the function of the protein. The multiple sequence alignment also showed good matches between gene 2505168929 and the cyanobacteria genes.

**Phospho-N-acetylmuramoyl-pentapeptide-transferase**

Gene OID: 2505169831EC: 2.7.8.13

Gene is predicted to code for the protein with EC 2.7.8.13. A NCBI BLAST search [11] gave phospho-N-acetylmuramoyl-pentapeptide-transferases as hits with e-values as low as 3E-129 to e-values as high as 2E-52, and around 60% identity. A PDB BLAST search was done, but PDB gave no results for EC 2.7.8.13. Despite the lack of a PDB result, the NCBI BLAST hits were strong, and they all were for phospho-N-acetylmuramoyl-pentapeptide-transferases, so the gene is predicted to code for EC 2.7.8.13.

**Undecaprenyldiphospho-muramoylpentapeptide beta-N-acetylglucosaminyltransferase**

Gene OID: 2505167482 EC: 2.4.1.227

Gene is predicted to code for EC 2.4.1.227. A NCBI BLAST search [13] gave UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferases as top hits, with low e-values (around 1E-100) and high percent identity (around 65%). A PDB BLAST search [7] gave an article on *E. coli* MurG with 35% identity and an e-value of 5.7086E-42. An analysis was made to see if the active site residues mentioned in the article matched with the ones in the *Microcoleus vaginatus* gene. 13 out of the 15 active sites matched. The mismatches were G102 to T102 and N135 to T135. The article did not mention anything about sites 102 nor 135. A multiple sequence alignment was done on T-Coffee with reviewed cyanobacteria FASTA sequences taken from UniProt. The results showed that at site 102, the T102 residue was almost completely conserved among cyanobacteria, with only one cyanobacteria having a S102 residue. At site 135, the T135 residue was completely conserved among cyanobacteria. This further confirms that gene 2505167482 codes for EC 2.4.1.227.

**Peptidoglycan glycosyltransferase**

Gene OID: 2505166662, 2505169221, 2505170013 and 2505170159 EC: 2.4.1.129

There were two genes predicted to code for peptidoglycan glycosyltransferase, also known as penicillin-binding protein. A NCBI BLAST search gave results on penicillin-binding proteins [22] with e-values as low as 2E-65 and as high as 6E-29, and around 30% identity. This BLAST also gave another set of results on peptidoglycan glycosyltransferases [25], after the first set of penicillin-binding proteins, with e-values of around 1E-20 and around 40% identity. A PDB BLAST search was also done, and an article on class A penicillin-binding proteins [16], with an e-value of 1.47901E-16 and 29% identity. The article mentioned 3 amino acid motifs that were conserved for such proteins. These motifs were SXXK, SXN and KTG, where the X’s stand for any amino acid. In the subject protein, the motifs were specifically STTK, SWN and KTG. For the query protein, a STFK motif was found, a SEN motif was found, and a KTG motif was found, so this makes it likely that the query protein indeed codes for EC 2.4.1.129. Finally, a reverse BLAST search was done with the PDB FASTA sequence, but this search gave two other genes as better hits before gene 2505166662. The found genes were 2505169221 and 2505170013, so these genes were checked as well.

A NCBI BLAST search was done with gene 2505169221, and this search also gave two sets of results. The first set of results were penicillin-binding protein hits [20] with e-values from 6E-99 to 3E-40 and around 31% to 37% identity. The second set of results were peptidoglycan glycosyltransferase hits [27] with e-values from 1E-31 to 6E-23 and 37% to 45% identity. A PDB BLAST search was also done, and the same PDB article [16] was a hit, with an e-value of 3.00846E-24 and 26% identity. The 3 conserved amino acid motifs were also found for this query protein, with the sequences being SSFK, SLN and KTG. This gene was the top hit for the reverse BLAST search.

A NCBI BLAST search was done with gene 2505170013, and this search gave two sets of results. The first set of results were penicillin binding protein hits [20] with e-values from 3E-90 to 8E-45 and 29% to 37% identity. The second set of results were petidoglycan glycosyltransferase hits [26] with e-values from 5E-31 to 6E-16 and 31% to 40% identity. A PDB BLAST search was done and also gave the same PDB article [16] as a hit, with an e-value of 7.32662E-22 and 26% identity. The same 3 conserved amino acid motifs were found in this query protein, with the sequences being STFK, SVN and KTG. This gene was the second hit for the reverse BLAST search.

Finally, gene 2505170159 was analyzed. This was the other gene predicted by IMG/ER to code for EC 2.4.1.129. A NCBI BLAST search was done and gave as top hits peptidoglycan synthase ftsI homologs [15] with e-valuse from 7E-130 to 3E-88 and 34% to 42% identity. A PDB BLAST search for articles corresponding to EC 2.4.1.129 gave the same PDB article [16] from before as a hit, with an e-value of 1.0884E-4 and 21% identity. The 3 conserved amino acid motifs were looked for again, and they were found to be STFK and SSN. The third KTG sequence was found, but not within the PDB multiple sequence alignment, since the KTG sequence was outside the frame of the multiple sequence alignment. A reverse BLAST search did not give this gene as a hit, but only the previously mentioned 3 genes as hits.

From this data, it seems that all four of the genes could code for EC 2.4.1.129. Further research needs to be done to see if only some of these genes or if all of these genes code for the protein, although e-values to the PDB hit and reverse BLAST search results suggest that gene 2505170159 is the least likely candidate for the protein.

**UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase**

Gene OID: 2505168660 EC: 6.3.2.10

Gene is predicted to code for EC 6.3.2.10. A NCBI BLAST search [18] gave as top results other UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligases, with low e-values (around 1e-60) and decent percent identity (around 30%). A PDB BLAST search [19] gave an artcile on UDP-N-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase, with 37% identity and an e-value of 1.62626E-60. This article mentioned a conserved sequence from sites 104 to 112 consisting of a Ala-Leu-Thr-Gly-Ser-Ser-Gly-Lys-Thr motif. This motif was somewhat conserved compared to gene 2505168660. The mismatches were L105 to V105 and S109 to V109. A multiple sequence alignment was done on T-Coffee, using cyanobacteria sequences from UniProt. At site 104, no other cyanobacteria had L105, and around half of them had V105. Additionally, leucine and valine are very similar in structure. At site 109, no other cyanobacteria had S109, and over half of them had V109. These were reasonable explanations for the active sites not matching, and gene 2505168660 is predicted to code for EC 6.3.2.10.

**UDP-N-acetylglucosamine 1-carboxyvinyltransferase**

Gene OID: 2505170551 EC: 2.5.1.7

This gene is predicted to code for EC 2.5.1.7, UDP-N-acetylglucosamine 1-carboxyvinyltransferase due to strong homology with other predicted 2.5.1.7 enzymes in PDB (PDB: 3SG1) and Swissprot databases (e-value 0.0, 73% identity) [12]. However, because none of the homologs were experimentally verified it may be necessary to do further research.

**Glutamate racemase**

Gene OID: 2505167410 EC: 5.1.1.3

This gene might code for Glutamate Racemase (EC 5.1.1.3). An NCBI BLAST search against the Swissprot database gave other Glutamate Racemase enzymes as top hits, with a query coverage of 94% and an e-value of 6e-106. [12] The PDB BLAST search gave Glutamate Racemase enzyme structures as top hits but none of the active sites matched. The top PDB hit had a 40% identity with an e-value of 1.58087e-55. [17] A reverse NCBI BLAST search also gave gene 2505167410 as a top hit, with an e-value of 0.

**UDP-N-acetylenolpyruvoylglucosamine reductase**

Gene OID: 2505167741 EC: 1.1.1.158

This gene is predicted to code for EC 1.1.1.158 due to strong homology with known UDP-N-acetylenolpyruvoylglucosamine reductase enzymes and close active site matching (10/12; mismatches included N80-A80 and Y77-L77). Experimentally verified homologs with EC 1.1.1.158 were found in Swissprot [5] and PDB databases (E-value 1.59E-53, 36% identity) [1].

**UDP-N-acetylmuramic acid:L-alanine ligase (MurC Synthase)**

Gene OID: 2505167742 EC: 6.3.2.8

This gene is predicted to code for EC 6.3.2.8 due to strong homology with known MurC Synthase enzymes and 100% active site conservation. An experimentally verified homology with EC 6.3.2.8 was found in PDB (33% identity, E-value 3.48E-52) [21].

**UDP-N-acetylmuramyl-tripeptide synthetase**

Gene OID: 2505168544 EC: 6.3.2.13

This gene is predicted to code for the UDP-N-acetylmuramyl-tripeptide synthetase enzyme (EC: 6.3.2.13). An NCBI BLAST search against the Swissprot database gave other UDP-N-acetylmuramyl-tripeptide synthetase enzymes as top hits, with a query coverage of 99% and an e-value of 0. [4] The PDB BLAST search gave the UDP-N-acetylmuramoyl-L-alanyl-D-glutamate enzyme structure as a top hit and almost all of the active sites matched. The top PDB hit had a 38% identity with an e-value of 1.12213e-73. [2]. A reverse NCBI BLAST search also gave gene as a top hit, with an e-value of 0.

**D-alanyl-D-alanine carboxypeptidase, serine-type**

Gene OID: 2505168109 EC: 3.4.16.4

This gene is predicted to code for the EC 3.4.16.4 enzyme due to strong homology with other experimentally verified DD-peptidases and 100% active site matching. Homologs were found in PDB (E-value 3.34932E-49, 31% identity) [24] and Swissprot databases

(E-value 7e-52, 33% identity) [23]. IMG predicted this enzyme to also have EC 3.4.99.- (D-alanyl-D-alanine-endopeptidase); however, none of the close homologs possessed that function. If this distinction is necessary to make, further annotation may be useful.

Gene OID: 2505166090 EC: ?

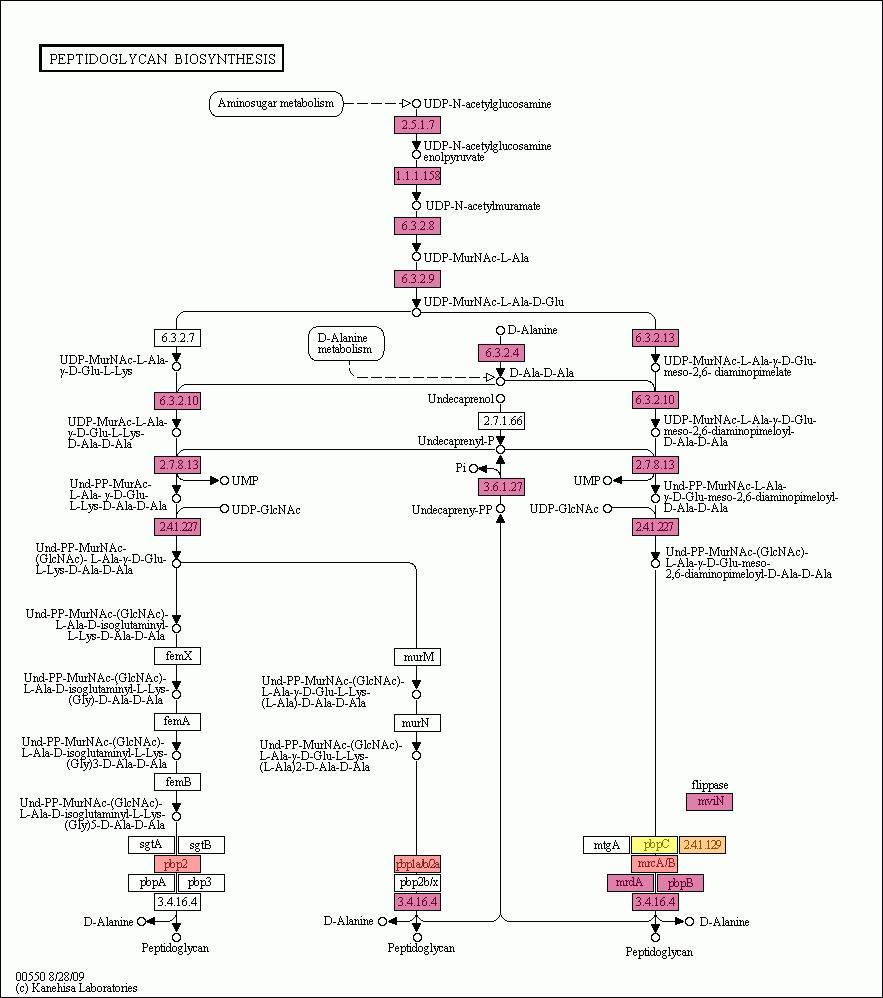
Although IMG predicted this gene to code for DD peptidase behavior it possessed weak homology with experimentally verified enzymes with EC 3.4.16.4. The top hit in PDB had an E-value ~3. A Swissprot BLAST returned similar weak homologs in addition to a suspected extracellular protein. In fact, many cyanobacteria seem to predict this gene, but also have weak homology with EC 3.4.16.4 enzymes. Becuase genome based predictions cannot fully predict the extraceullar proteome [28] further annotation/research may be necessary to characterise this gene.

**D-alanyl-D-alanine carboxypeptidase (penicillin-binding protein 4)**

Gene OID: 2505168263 & 2505169683 EC: 3.4.16.4/3.4.99.-

These genes are predicted to code for the EC 3.4.16.4/3.4.99.- enzyme due to homology with other experimentally verified DD-peptidases and close active site matching. Homologs for both were found in PDB (E-value 2.16E-08/4.53E-10, 28%/36% identity) [3] and Swissprot databases (E-value 2e-07/2e-09 28%/36% identity) [14]. The active site S62 is replaced by Q62 but this is a minor change such that this enzyme may still be able to function properly. No strong homologs were found because enzymes with EC 3.4.99.- are not completely understood.

**Comparison with Other Genomes**





S Fig. 2: Comparison of genes found in *Microcoleus vaginatus* with other cyanobacteria in IMG/ER for peptidoglycan biosynthesis

*Microcoleus vaginatus* follows the right branch of S Figure 2, an image from Kegg. This corresponds to the pathway shown from S Figure 1 from MetaCyc. Most of the genes found in *Microcoleus vaginatus* necessary for the pathway were found in more than 75% of other cyanobacteria. The only anomaly was the gene coding for EC 2.4.1.129,which was predicted to only be in around 25% to 50% of cyanobacteria. This analysis shows though that *Microcoleus vaginatus* has genes that code for proteins that are documented in other cyanobacteria.

**Functional Studies**

Functional studies should be done on genes 2505166662, 2505169221, 2505170013 and 2505170159, which all could potentially code for EC 2.4.1.129. Experiments need to done to show which genes actually code for the protein. Experiments could be done with different combinations of the genes repressed, and it can be observed whether the protein is still expressed or not, so that conclusions can be made about which genes are linked to protein expression.

Gene 2505167410 was another subject for functional studies. This gene was predicted to code for glutamate racemase, but none of the active site residues matched with those mentioned in the literature. The PDB, NCBI and reverse BLAST searches all gave good results though, so it is unclear from gene annotation whether or not this gene actually codes for glutamate racemase. An experiment needs to be done where the gene is repressed, and it should be seen if glutamate racemase is still expressed or not.

Gene 2505166090 was predicted via IMG to code for an enzyme with EC 3.4.16.4 yet it possessed very weak homology with known DD-peptidases. Because similar genes were found in many other cyanobacteria, it may be useful to further characterize this protein with a lab experiment. Also, experiments can be done on gene 2505168109 because this gene is predicted to code for a bifunctional protein with function for EC 3.4.16.4 and EC 3.4.99.-. An experiment should be done to see if this gene is actually bifunctional. The gene should be repressed and it should be seen if any proteins with function for EC 3.4.99.- are still expressed, assuming no other genes code for proteins with EC 3.4.99.- function.

**Discussion**

All the genes needed for peptidoglycan biosynthesis were found and annotated. However, some genes were problematic in that they did not clearly code for a certain protein. Genes 2505166662, 2505169221, 2505170013 and 2505170159 all potentially code for EC 2.4.1.129. Gene 2505167410 was predicted to code for glutamate racemase, but further studies need to be done to confirm this because none of the active sites matched with the ones mentioned in the literature. Genes 2505166090 and 2505168109 were predicted to code for EC 3.4.16.4 and for EC 3.4.16.4 and 3.3.99.-, respectively. The first gene needed to be checked to see whether it codes for EC 3.4.16.4, while the second gene needs to be checked to see if it codes for a bifunctional protein.

Overall, this pathway is predicted to be complete and should be complete for *Microvoleus vaginatus* because the organism makes cell walls. This annotation merely confirms common sense intuition that the organism has a mechanism to create its own cell walls.

**References**

[1] Benson TE, Harris MS, Choi GH, Cialdella JI, Herberg JT, Martin JP Jr, Baldwin ET. "A

structural variation for MurB: X-ray crystal structure of Staphylococcus aureus

UDP-N-acetylenolpyruvylglucosamine reductase (MurB)." Biochemistry. Feb 27, 2001;

40(8):2340-50.

[2] Gordon E, Flouret B, Chantalat L, van Heijenoort J, Mengin-Lecreulx D, Dideberg O. "Crystal

structure of UDP-N-acetylmuramoyl-L-alanyl-D-glutamate: meso-diaminopimelate ligase

from Escherichia coli." J Biol Chem. Apr 6, 2001; 276(14):10999-1006.

[3] Kamtekar S, Berman AJ, Wang J, Lázaro JM, de Vega M, Blanco L, Salas M, Steitz TA. "The

phi29 DNA polymerase:protein-primer structure suggests a model for the initiation to

elongation transition." EMBO J. Mar 22, 2006; 25(6):1335-43.

[4] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[5] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[6] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[7] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[8] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[9] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

10] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[11] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[12] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[13] Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa

A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A,

Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S.

"Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium

Anabaena sp. strain PCC 7120." DNA Res. Oct 31, 2001; 8(5):205-13; 227-53.

[14] Korat B, Mottl H, Keck W. "Penicillin-binding protein 4 of Escherichia coli: molecular cloning

of the dacB gene, controlled overexpression, and alterations in murein composition." Mol

Microbiol. 1991 Mar; 5(3):675-84.

[15] Lemieux C, Otis C, Turmel M. "Ancestral chloroplast genome in Mesostigma viride reveals

an early branch of green plant evolution." Nature. Feb 10, 2000; 403(6770):649-52.

[16] Lemieux C, Otis C, Turmel M. "Ancestral chloroplast genome in Mesostigma viride reveals

an early branch of green plant evolution." Nature. Feb 10, 2000; 403(6770):649-52.

[17] Lundqvist T, Fisher SL, Kern G, Folmer RH, Xue Y, Newton DT, Keating TA, Alm RA, de

Jonge BL. "Exploitation of structural and regulatory diversity in glutamate racemases."

Nature. Jun 14, 2007; 447(7146):817-22.

[18] Malakhov MP, Los DA, Wada H, Semenenko VE, Murata N. "Characterization of the murF

gene of the cyanobacterium Synechocystis sp. PCC 6803." Microbiology. 1995 Jan; 141 (Pt 1)():163-9.

[19] Malakhov MP, Los DA, Wada H, Semenenko VE, Murata N. "Characterization of the murF

gene of the cyanobacterium Synechocystis sp. PCC 6803." Microbiology. 1995 Jan; 141 (Pt 1)():163-9.

[20] McPherson DC, Driks A, Popham DL. "Two class A high-molecular-weight penicillin-binding

proteins of Bacillus subtilis play redundant roles in sporulation." J Bacteriol. 2001 Oct;

183(20):6046-53.

[21] Mol CD, Brooun A, Dougan DR, Hilgers MT, Tari LW, Wijnands RA, Knuth MW, McRee DE,

Swanson RV. "Crystal structures of active fully assembled substrate- and product-bound

complexes of UDP-N-acetylmuramic acid:L-alanine ligase (MurC) from Haemophilus

influenzae." J Bacteriol. 2003 Jul; 185(14):4152-62.

[22] Popham DL, Setlow P. "Cloning, nucleotide sequence, and regulation of the Bacillus subtilis

pbpF gene, which codes for a putative class A high-molecular-weight penicillin-binding

protein." J Bacteriol. 1993 Aug; 175(15):4870-6.

[23] Sauvage E, Herman R, Petrella S, Duez C, Bouillenne F, Frère JM, Charlier P. "Crystal

structure of the Actinomadura R39 DD-peptidase reveals new domains in

penicillin-binding proteins." J Biol Chem. Sep 2, 2005; 280(35):31249-56.

[24] Sauvage E, Herman R, Petrella S, Duez C, Bouillenne F, Frère JM, Charlier P. "Crystal

structure of the Actinomadura R39 DD-peptidase reveals new domains in

penicillin-binding proteins." J Biol Chem. Sep 2, 2005; 280(35):31249-56.

[25] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[26] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[27] Schmoldt A, Benthe HF, Haberland G. "Digitoxin metabolism by rat liver microsomes."

Biochem Pharmacol. Sep 1, 1975; 24(17):1639-41.

[28] Tjalsma H, Antelmann H, Jongbloed JD, Braun PG, Darmon E, Dorenbos R, Dubois JY,

Westers H, Zanen G, Quax WJ, Kuipers OP, Bron S, Hecker M, van Dijl JM. "Proteomics of protein secretion by Bacillus subtilis: separating the "secrets" of the secretome." Microbiol Mol Biol Rev. 2004 Jun; 68(2):207-33.