Proposed Regulatory sequence that influences Transcription of the Carboxylase Gene in Prochlorocus and Syneococus

EJ Kochis

**Introduction**

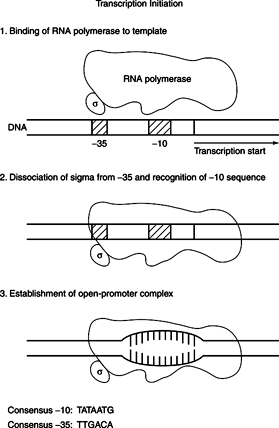
The central dogma of biology is that DNA is transcribed into RNA and RNA is translated into proteins. When DNA is transcribed in in bacteria, there is a promoter region that the RNA polymerase binds to, in order to begin transcription. Promoter regions in bacteria have two distinct sequences that are recognizable by the RNA polymerase complex, the -35 and the -10. These numbers, denote how many nucleotides upstream from the transcription start site that the sequences are in. There exists a consensus promoter for the -35 and -10 regions (Hawley and McClure 1983). These consensus sequences are the nucleotides that are most likely to occur in a bacterial promoter. The consensus sequence for bacteria the -35 region is TTGACA and TATAATG for -10. Figure 1 displays the general process of transcription in bacteria along with the consensus promoter.

Figure 1. Displays a general outline of how transcription occurs in bacteria. The -35 and -10 consensus sequences are also shown

Although a consensus sequence does exist, not all promoters will match the consensus, which makes finding functional promoters especially difficult to do computationally. Despite this setback, promoters are still heavily studied. In one study done by , Vogel, et al in 2003 researchers found a conserved potential promoter in the upstream sequence of a Carboxylase gene. This conserved promoter was only examined in 4 organisms, two species of Prochlrococcus and two species of Synenococcus. The conserved promoter is below in Figure 2. The -35 region highlighted is fully conserved, while the -10 region is less conserved and the transcription start site or TSS is also fully conserved. The purpose of this experiment is to extend Vogel’s work and attempt to confirm Vogels results that the sequence is a promoter; I also want to find find the conserved Carboxylase sequence in many different species of Prochlrococcus and Synenococcus that have not yet been studied.

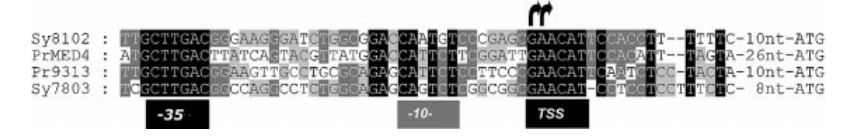


Figure 2. Shows the original sequences found by Vogel et al in Synechococcus WH 8102/7803 and Prochlrococcus MED4/MIT9313

**Method**

I started my experiment by first confirming Vogel et al’s results by finding the conserved sequences that the researchers found. To do this I gathered the genomes of Synechococcus WH8102/7803 and Prochlrococcus MED4/MIT9313. The full genomes can be found using National Center for Biotechnology Information search engine, the genbank accession numbers for the four organisms are:

Synechococcus WH 8102: BX548020.1

Synechococcus WH 7803: NC\_009481.1

Prochlrococcus MED4: BX548174.1

Prochlrococcus MIT9313: BX548175.1

Once the full genomes of the organisms were found I used a tool called Cyano biobike to analyze the sequences and find the potential promoter that Vogel et al found. To find that promoter that I used a function in Cyanobike called “MATCHES-OF-PATTERN”, this function takes an input and attempts to find a match of that input in a genome, a visual representation of the function and the results it returned can be found in Figure 3. Because by nature promoters are found outside of genes I specified my function to only look in the intergenic regions of all the organisms. When I ran the function I was mainly just looking for the two fully conserved sequences which is the -35 GCTTGAC and the transcription start site GAACAT outlined in figure 2, this is why the syntax in the function is GCTTGAC.\*GAACAT, this notation with the “.” And the “\*” tells the function to find the two sequences I specified with anything in-between them because that area is not as conserved.





Figure 3. This shows the function I used in Cyanobike along with the result which tells me that the sequence is found in the 889th intergenic region from coordinates 68-22 within that intergenic sequence.

The first step in finding the conserved sequence was to gather a list of organisms that the potential promoter could be in. To do this, I compiled a list of 10 species of Prochlrococcus and 14 species of Synechococcus. The full list of organisms can be found in the supplements. Using Cyanobike I was able to gather the sequenced genomes of the species in question. In order to find the potential carboxylase promoter I ran the same function as in figure 3 but with the genomes of the other 21 cyanobacteria that I had access to.

If the sequence was found in an organism I used a “CONTEXT-OF” function to check that the sequence was in-fact upstream of carboxylase and not promoter for a different gene. The context of function showed me a visual representation where my potential promoter was in relation to the genes around it, below is a picture of what the function returned, the purple arrow is my promoter and the red gene is a Carboxylase gene. Every potential promoter found was indeed upstream of a Carboxylase gene, confirming my preliminary results.

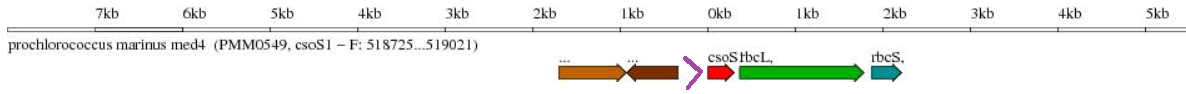
**RESULTS**

Figure 4 This what CONTEXT-OF function returns. The purple arrow is the promoter that was found and the genes to the right are carboxylase genes that are being transcribed.

After all 24 organisms were checked for the promoter, there were 12 instances found. Of the 12 matches of the potential promoter found, there were 6 in Prochlrococcuses and 6 in Synechococcus. A full list of the sequences found is below, the -35 and the TSS sites are highlighted to show the conserved sequence. They are in the same notation shown in figure 3 as described in the methods section\*.

\*Due to Cyanobike errors, some of the results returned raw coordinates such as Synechococcus Bl107 which returned coordinates, 1068431 1068475 these are the coordinates that spanned the sequence.

Prochlrococcus potential Carboxylase promoters:

Prochlororcocus Marinius as9601

312 259 303 GCTTGACTTATAAGTACTTAATGAAGCATTCTTCGGATTGAACAT

MED4

296 258 302 GCTTGACTTATCAGTACGTTATGGACCATTCTTCGGATTGAACAT

MIT 9301

299 262 306 GCTTGACTTATAAGTACTTAATGAAGCATTCTTCGAATTGAACAT

MIT 9303

467 99 143 GCTTGACGGAAGGATGGGCTGCAGAGCATTGTCCATTGCGAACAT

MIT 9313

823 69 23 GCTTGACGGAAGTTGCCTGCGCAGAGCATTCTCCTTCCCGAACAT

MIT 9515

335 259 303 GCTTGACTTATAAGTACGTTATGGACCATTCTTCCAATTGAACAT

Synechococcus Carboxylase promoters:

Bl107\*

1068431 1068475 GCTTGACGGCGAGGGTCTTGGCAGAGCAATGTCCCGAGCGAACAT

CC9311

1173 69 23 GCTTGACGAGTTGGCCTCCGGCAGACCAATGTCATCGGCGAACAT

CC9605

399 262 306 GCTTGACGGGAAGGGTTCTGGCGAAGCAATGTCCCCAGCGAACAT

CC9902

900 68 22 GCTTGACGGCGAGGGTCTTGGCAGAGCAATGTCCCGAGCGAACAT

WH 7803

348 223 267 GCTTGACGGCCAGGCCTCTGGCAGAGCAGTCTCGGCGGCGAACAT

WH 8102

880 68 22 GCTTGACGGGAAGGGATCTGGCGGACCAATGTCCCGAGCGAACAT

These were the raw sequences that were found using the matches of pattern function. Once I got these sequences I set out to make a consensus potential promoter sequence from them. A consensus sequence is a sequence that takes the most frequent nucleotide at each position from a number of sequences and makes a consensus from them. In order to form my consensus sequence I used Cyanobike and formed an alignment of the promoters using an ALIGNMENT-OF function, the results are below in figure 5. The 45 nucleotide consensus sequence that was returned is:

GCTTGACGGATATGGGA [T/C] [T/G] TGGC [A/G] GAGCATT [G/C] TCCCGAGCGAACAT

**-35 -10 TSS**

The highlighted areas showed 100% conservation throughout all the seqeunces, as expected -35, -10 and TSS were the most highly conserved areas. The positions that have two nucleotides in them such as T/C means that no nucleotide occurred in high enough frequency to form a consensus but those two nucleotides are the most likely at those positions.

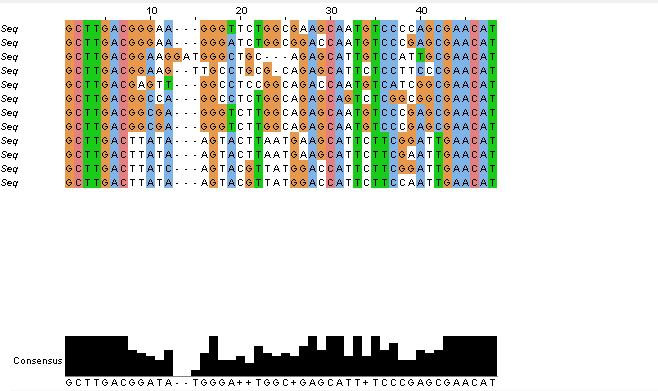


Figure 5 This figure shows the alignment and subsequent consensus sequence of the 12 promoters that were found in the organisms.

**Statistical Significance**

In order to test the rarity and possible statistical significance of this promoter I needed to first look at the nucleotide content of the intergenic regions of the organisms used in this experiment to get a reference to search for the sequence. I used CyanoBike to get the average GC-fraction of all the intergenic sequences of the 12 organisms that the promoter was found in. The result returned an average GC-fraction of .40. I then needed to make a sample of random DNA to check if the sample is found randomly, using Cyanobike I made set of random DNA 10,000,000 nucleotides long with a GC content of .40. I then checked for a match of the fully conserved -35 region and the TSS, I repeated the function 20 times, meaning I looked for the sequence in 200 million nucleotides. I got 0 viable matches; this means that it is highly unlikely that the sequence is due to chance. Another method of finding statistical significance would be to use the Poisson expression. Poisson expression is a statistical test of expected results versus observed results. In order to test the amount of expected matches of the promoter I gathered the entire length of all the intergenic sequences that the promoter was found in which is 14,988\* then I calculated to probability of the -35 and TSS site occurring using the GC fraction for the frequencies of the nucleotides. I found that by chance the -35 and TSS sites would be found once every 71,444,901 nucleotides, this number was obtained from one over the probability of the -35 and TSS sites P=1.399e-8

Frequency = 1/1.399e-8= 71444901

Using this expected frequency we would expect 14988/71444901=2.09e-4 hits

Then I imputed my expected number **u = 2.09e-4** and my **actual hits x=12** into a poisson expression

**P(*x*; μ) = (e-μ) (μx) / x!**

**P(12;2.09e-4)= (e^-2.09e4) (2.09e-4^12) / 12!**

**The result of this expression is equal to P= 1.44986545290769E-53**

This number is astronomically low which means that it is extremely unlikely that the potential promoter is due to chance.

\*The Cyanobike functions used in this section can be found in the appendix

**Discussion**

Based upon my results above I expect that Vogel et al did indeed find a promoter for Carboxylase. There is high conservation in the -35 and TSS regions of the sequences for many Prochlrococcus and Synechococcus. Also, the -35 region of the Carboxylase promoter is very similar to the bacterial consensus promoter which has proved functional in many organisms (Sutdnicka).

🡨-35 region for Experimental carboxylase promoter

sd

GCTTGAC

🡨-35 region for confirmed bacterial consensus sequence

TTGACA

Given the consensus sequence of the carboxylase potential promoter I was able to use the BLAST tool on NCBI to run that consensus sequence against known cyanobacteria, below are the results of that blast search (Figure 6). I got 200 matches to the consensus sequence. Using this list one could do a similar experiment and find out if the promoter is functional for carboxylase in these genomes as well.

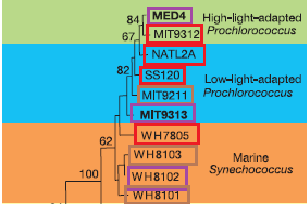
After examining the phylogenetic tree based off the 16-23s rRNA of Prochlrococcus and Synechococcus the results can be seen from a different angle. Figure 7 shows a tree of Prochlrococcus and Synechococcus the tree displays how similar the sequences of each organism are to each other. The figure also shows if I found the promoter the species is highlighted in purple, if I did not find the sequence then it is red, and if I did not look in the genome then the organism is highlighted brown. There are 3 cases of the promoter in the tree, one in each ‘clade’.

Figure 7 1 The Tree shows how related the genomes of each organism is to each other, the numbers on the left denote a percent of similarity.

One of the most interesting points that I came across in literature is that all cyanobacteria discovered thus far have the micro-organelle Carboxylase (Kinney et al). Yet according to my findings, the Carboxylase promoter is obviously not 100% conserved throughout all of the cyanobacteria on earth. Further research could be done on why some of the promoters are conserved and others are not, is my consensus promoter not as effective? Or could I be more effective but just not used by other organisms? These are experiments that could be in a lab setting, which could be difficult to execute computationally.

Overall I was able to extend Vogel’s original research and find 8 more cases of the Carboxylase promoter, and form a consensus promoter that could be used in future research to find more cases occurrences in more cyanobacteria. My pool of organisms what somewhat small compared to total known cyanobacteria, I have no doubt that there are more cases of this promoter I was just unable to analyze them in this study.

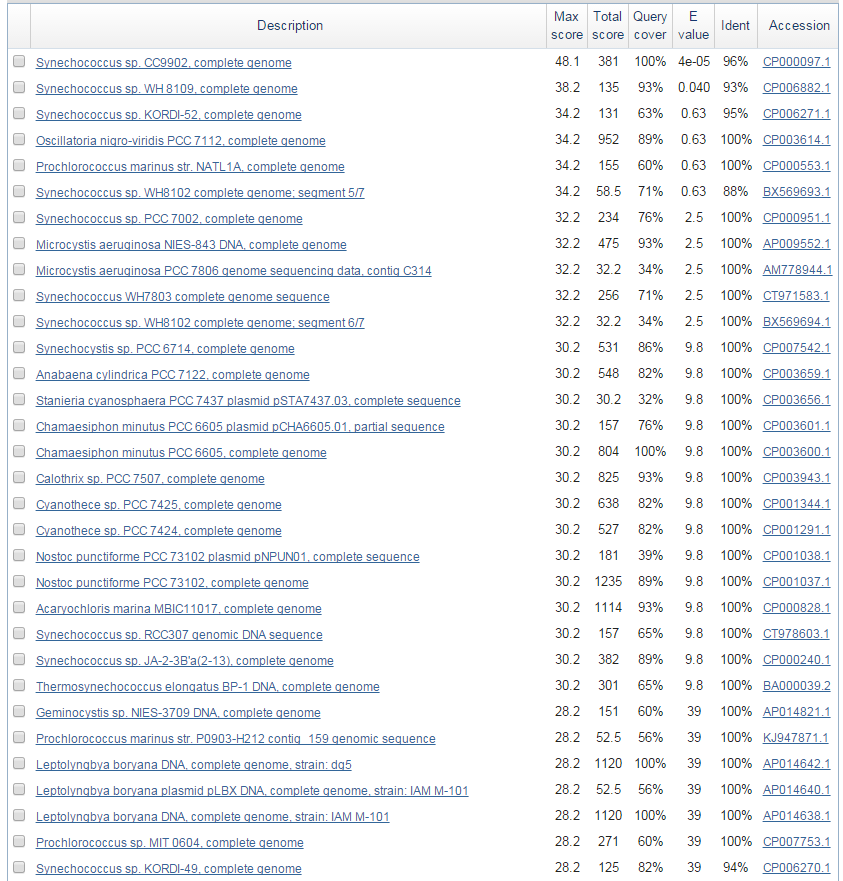


Figure 6 This figure displays the matches of the BLAST search in NCBI. This list of organisms could be used to extend this research

**References**

Elhai J, Taton A, Massar JP, Myers JK, Travers M, Casey J, Slupesky M, Shrager J (2009). BioBIKE: A web-based, programmable, integrated biological knowledge base. Nucl Acids Res 37:W28-W32

Hawley, D. K. & McClure, W. R. (1983) Nucleic Acids Res. 11, 2237-2255

Gary M. STUDNICKA. "Escherichia Coli Promoter -10 and -35 Region Homologies Correlate with Binding and Isomerization Kinetics."*Biochem.* (1988) 252, 825-831 (n.d.): n. pag. Web. 14 Apr. 2015.

Kinney, James N., et al. "Elucidating essential role of conserved carboxysomal protein CcmN reveals common feature of bacterial microcompartment assembly." *Journal of Biological Chemistry* 287.21 (2012): 17729-17736.

"Transcription in Prokaryotes." *Transcription in Prokaryotes*. CliffsNotes, n.d. Web. 27 Apr. 2015.

Vogel, Jörg, et al. "Experimental and computational analysis of transcriptional start sites in the cyanobacterium Prochlorococcus MED4." *Nucleic acids research* 31.11 (2003): 2890-2899.

**Appendix**

**Raw list of all organisms searched and promoter sequence that was found and if it was not found then it says ‘not found’**

List organisms:

Prochlororcocus Marinius as9601

312 259 303 GCTTGACTTATAAGTACTTAATGAAGCATTCTTCGGATTGAACAT

MED4

296 258 302 GCTTGACTTATCAGTACGTTATGGACCATTCTTCGGATTGAACAT

MIT9211

NOT FOUND

MIT 9301

299 262 306 GCTTGACTTATAAGTACTTAATGAAGCATTCTTCGAATTGAACAT F

MIT 9303

467 99 143 GCTTGACGGAAGGATGGGCTGCAGAGCATTGTCCATTGCGAACAT

Mit 9312

NOT FOUND

MIT 9313

823 69 23 GCTTGACGGAAGTTGCCTGCGCAGAGCATTCTCCTTCCCGAACAT B

MIT 9515

335 259 303 GCTTGACTTATAAGTACGTTATGGACCATTCTTCCAATTGAACAT

Nat1a

NOT FOUND

Nat2a

NOT found

SS120

NOT FOUND

SYNE

Bl107

1068431 1068475 GCTTGACGGCGAGGGTCTTGGCAGAGCAATGTCCCGAGCGAACAT

CC9311

1173 69 23 GCTTGACGAGTTGGCCTCCGGCAGACCAATGTCATCGGCGAACAT B

CC9605

399 262 306 GCTTGACGGGAAGGGTTCTGGCGAAGCAATGTCCCCAGCGAACAT

CC9902

900 68 22 GCTTGACGGCGAGGGTCTTGGCAGAGCAATGTCCCGAGCGAACAT

Elgotnat

NOT FOUND

Pcc 7002

Not found

Pcc 7336

Not found

Rcc 307 Not found

Rs9916

Not found

Rss9917

Not found

WH 5701

Not Found

WH 7803

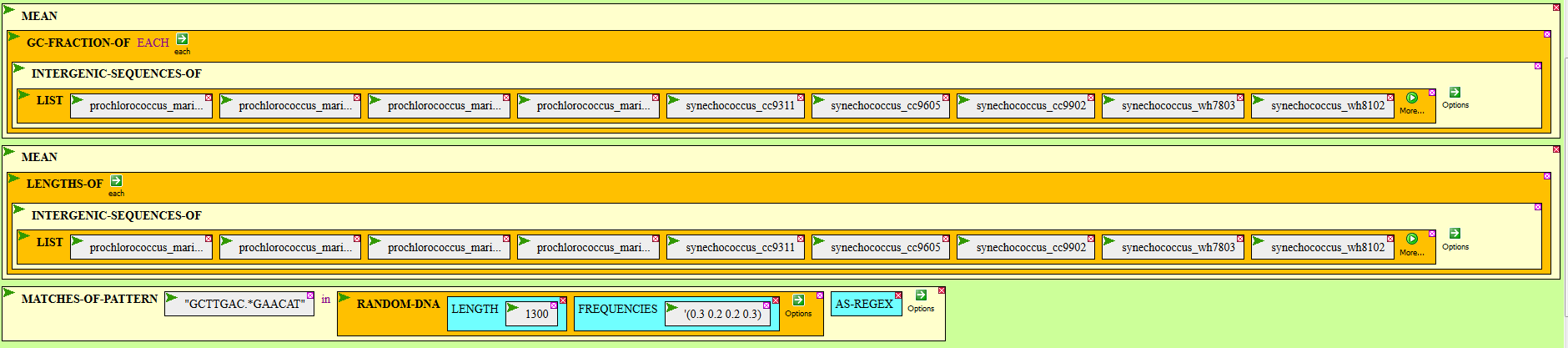
348 223 267 GCTTGACGGCCAGGCCTCTGGCAGAGCAGTCTCGGCGGCGAACAT F

WH 7805

Not found

WH 8102

880 68 22 GCTTGACGGGAAGGGATCTGGCGGACCAATGTCCCGAGCGAACAT

**Below is the Cyanobike code that I used to get the mean GC-fractions and mean length of intergenic sequences**