Investigation in HIP1-Related Sequence in Cyanobacteria A and B

Michael Vuong
BNFO 301
Spring 2014

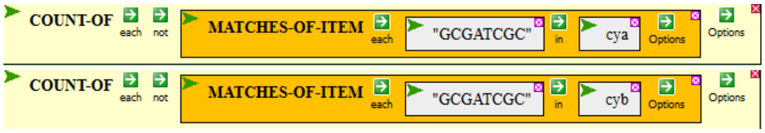
Introduction

Palindromes, in the context of genetics, are nucleotide sequences that are read the same 5' (prime) to 3' on one strand or 5' to 3' on the complementary strand. They typically play important roles in molecular biology such as acting as restriction enzyme sites. The palindrome that was being explored was an octameric (8-nucleotides long) palindrome known as highly iterated palindrome 1 (HIP1) with the sequence of 5'-GCGATCGC-3' (3'-CGCTAGCG-5' on complementary strand)[1]. This sequence was unique due to the fact that it was highly represented among many cyanobacteria genomes, meaning it has a very high occurrence rate. For example, the table in Figure 1 displayed the most common octameric sequence found in a given cyanobacteria, Synechococcus elongatus pcc6301. Not surprisingly, HIP1 was the most found sequence at 7356 matches in the genome. More interesting, however, was the fact that every sequence afterwards in terms of number of matches was simply HIP1 with one or two different nucleotide at the beginning or end of the sequence. For example, the second most found octameric sequence in the table, "GGCGATCG", was essentially HIP1 without the cytosine at the end of the sequence and a guanine added to the beginning of the sequence. This pattern continued for a large portion of the output, indicating a significant repeat of this sequence in Synechococcus elongatus pcc6301. Even without this fact, the HIP1 sequence occurred more than twice the amount of the second most found octameric sequence. This pattern occurred for multiple cyanobacteria, which suggests HIP1 should have a significant impact and role in these organisms.

According to the phylogenic tree in Figure 2, the boxed regions indicate cyanobacteria that do not have a large abundance of HIP1 in their genome, indicating that HIP1 was not overrepresented in every cyanobacteria. This was observed in marine picocyanobacteria in the bottom portion of the tree as well as in a group at the top of the tree containing gloeobacter violaceus, cyanobacteria A (CYA), and cyanobacteria B (CYB).   This begged the question, if HIP1 was so highly represented and presumably important to cyanobacteria, why do these groups of cyanobacteria not have it in similar abundance. There were many possible reasons for this such as these cyanobacteria do not require the function that HIP1 serves, they have a different sequence that serves the same purpose, HIP1 does not have a specific function that was crucial to all cyanobacteria and it was a coincidence they are highly represented, or there was an evolutionary change such that HIP1 developed for cyanobacteria to better adapt.

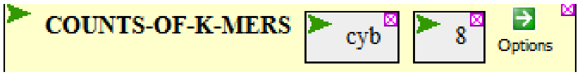
Methods

To identify what was unique about the cyanobacteria that do not appear to have an abundance of HIP1, specifically CYA and CYB, the first step was to confirm that both had very low HIP1 count. This was done through the use of COUNT-OF, MATCHES-OF-ITEM, and the sequence "GCGATCGC" for both CYA and CYB, resulting in 134 and 114 number of matches respectively. This result was much lower than the number of occurrences in the cyanobacteria near the center of the phylogenic tree in figure 2 where possible results were in the thousands such as 14712 matches for Synechococcus elongatus PCC6301 and 7362 matches for Thermosynechococcus elongatus BP1.



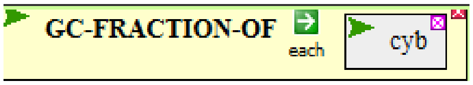
```
COUNT-OF   MATCHES-OF-ITEM   "GCGATCGC"   in   cya
COUNT-OF   MATCHES-OF-ITEM   "GCGATCGC"   in   cyb
```

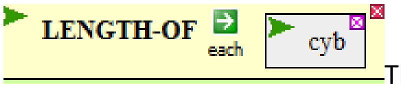
From this point, the theory that there was a replacement sequence was tested by running Counts-of-K-mers on CYA and CYB with window size 8 to search for the most common octameric sequence and determine if there was another sequence that could fit the role of HIP1 in these cyanobacteria. Interestingly, the results, shown in figure 3 and 4, indicate the most common sequence in both CYA and CYB are the identical: “GGGATCCC”.


COUNTS-OF-K-MERS cyb 8 Options

More interesting was the fact that this sequence was actually a palindrome and was very similar to the original HIP1 sequence with only a 2 nucleotide difference. Additionally, the next most common sequences follow the same pattern as the output from Synechococcus elongatus pcc6301 where they are various shifts of the most common sequence. Furthermore, the 3 most common sequences were identical for both organisms.

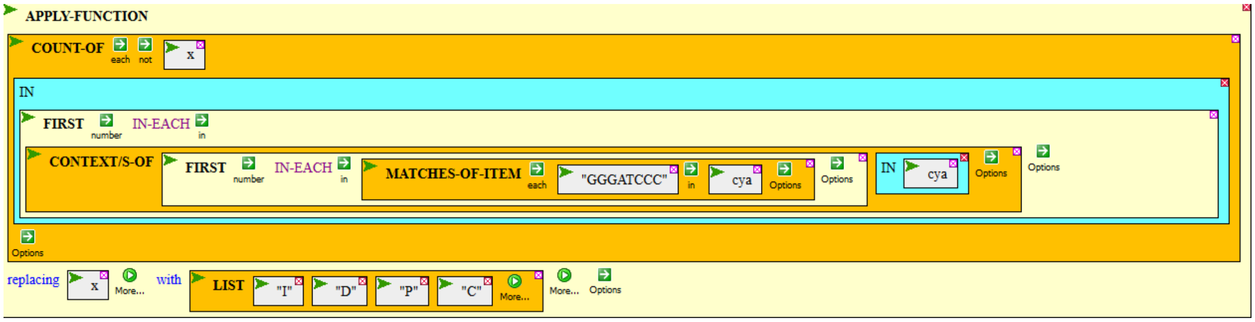
To test the significance of this sequence, the GC-FRACTION-OF and LENGTH-OF functions were used to calculate the probability of each nucleotide.


GC-FRACTION-OF each cyb


LENGTH-OF each cyb

The probability of nucleotides G or C occurring in the genome was GC fraction divided by 2 and the probability of nucleotides A or T occurring in the genome was (1- GC fraction)/2. These probabilities are used to determine the probability of the sequence occurring in the respective genomes. This probability was multiplied by the length of the genome to determine the expected number of occurrences. The results of the calculations are displayed in Table 1.

Additionally, the behavior of this sequence was observed in terms of location such as whether it was inside a gene (I), parallel (P), divergent (D), or convergent (C). Parallel indicates that the sequence is found between the end of one gene and beginning of another. Divergent indicates the sequence is found between the beginnings of two genes and convergent indicates the sequence is between the ends of two genes. This was done by searching for every instance the sequence is found in each organism using MATCHES-OF-ITEM and using CONTEXT-OF on the coordinates of each found sequence to identify genes nearby. Using FIRST-IN EACH, the first element of each context found was extracted to indicate whether the sequence was found inside or out of the gene by the output of four letters: I, P, D, C. Finally, APPLY-FUNCTION was used to loop through the count of the amount of times each letter appeared, indicating how often the sequence was in or out of coding regions. Since this function counted the occurrence of the sequence in both strands, the output was divided by two to obtain the accurate number of times the sequence was found. The results from this function is displayed in Table 2.



```
APPLY-FUNCTION
  COUNT-OF each not x
    IN
      FIRST number IN-EACH in
        CONTEXT/S-OF FIRST number IN-EACH in MATCHES-OF-ITEM each "GGGATCCC" in cya Options Options
                     IN cya Options
      Options
    Options
  replacing x More... with LIST "I" "D" "P" "C" More... More... Options
```

Results

Table 1 indicates the expected versus observed ratio of the found sequence “GGGATCCC” for both CYA and CYB. Interestingly, the ratios are essentially 39 and 50 times higher than expected for CYA and CYB respectively. This number is comparable to the ratios found by Delaye and Moya (2011) in organisms where the HIP1 sequence was highly represented.[1] Additionally, a search through MATCHES-OF-ITEM for this found sequence in an organism known to have high representation of HIP1, Synechococcus elongatus pcc6301, resulted in only 72 hits compared to 14712 hits of HIP1. Table 2 shows the results of the algorithm to identify where the sequence is mostly occurring. A large majority of the found sequence in both CYA and CYB are located in coding regions/within genes, 2825 and 3295 respectively. A total of 576 and 804 occurrences of the sequence for CYA and CYB respectively were found outside of the coding region, with the majority being parallel in both organisms.

Table 1: Expected vs Observed Occurrence of “GGGATCCC”

Organism	GC Content	Length	Probability of “GGGATCCC”	Expected Number of Occurrence	Observed-Expected Ratio
CYA	0.6023733 G=C=0.30118665 A=T=0.19881335	2932766	(0.30118665)^6 * (0.19881335)^2 =2.9505*10^-5	86.533	3401/86.533 =39.3029
CYB	0.5845034 G=C=0.2922517 A=T=0.2077483	3046682	(0.2922517)^6 * (0.2077483)^2 =2.68916*10^-5	81.9303	4099/81.9303 =50.03033

Table 2: Occurrence of “GGGATCCC” in Coding Vs. Non-coding Regions

Organism	Occurrence of Sequence In Gene	Occurrence of Sequence in Parallel	Occurrence of Sequence in Divergence	Occurrence of Sequence In Convergence
CYA	2825	315	195	66
CYB	3295	460	218	126

Discussion

Based on the results found, it is very possible that the sequence found is a mutation or prior form of the HIP1 sequence due to the similarity with only a two nucleotide difference. It was found that HIP1 occurs more often in coding regions than non-coding regions, which is similar to the characteristics of this sequence as shown in Table 2 [1]. More than 80 percent of the found sequences in both CYA and

CYB are located in coding regions, however, even among non-coding region occurrences, there was a bias towards divergent regions, meaning the more sequences occurred near the beginning of two genes when present in non-coding regions. Considering the large observed to expected ratio for both, 39.3029 and 50.03033, this sequence, even if unrelated to HIP1, is likely to have a(n) important role(s) in the two organisms. When doing the calculations, the expected number of the found sequence is identical to the expected number of HIP1 sequence since both sequences share the same number of GC and AT content, thus unable to distinguish between the two. However, the MATCHES-OF-ITEM of the HIP1 sequence in CYA and CYB, resulting in 134 and 114 matches, leads to a little greater than 1 observed to expected ratio. Additionally, when looking at the most found sequences in Figure 3 and 4, the pattern of shifts of the most common sequence is similar to figure 1 of the HIP1 sequence occurrence. This may suggest that some mutation occurred to this found sequence between the different groups of cyanobacteria from the phylogenic tree to make the found sequence turn into HIP1 as the most common sequence. HIP1 has been suggested to serve a role as a recognition site for DAM methylase, which attaches to "GATC" sequences to control DNA replication [1]. As the found sequence contains the recognition site for DAM methylase, it may also share that role if such a function exists.

Unfortunately, this is not sufficient information to confirm there is a relationship between HIP1 and the found sequence. Additionally, it is also unknown whether there is significance in the overwhelmingly high occurrence of the found sequence in coding regions in CYA and CYB compared to non-coding regions. Further testing would have to be done on HIP1 to clearly identify its function, if it exists, in cyanobacteria and determine if it can be applied to the found sequence. It may also be possible to search through other organisms, such as the multiple marine picocyanobacteria, where the observed versus expected ratio is also much lower than the other cyanobacteria. However, since these marine cyanobacteria are essentially a different group than the CYA and CYB tested, any findings may not be relevant to the findings in this investigation. As suggested earlier in this report, many reasons exist and should be tested for certain cyanobacteria not having an abundance of HIP1 may be a result of a different sequence, HIP1 having no function, or these cyanobacteria do not require whatever function HIP1 may serve.

Appendix

Figure 1.

```
COUNT     WORD
 7356    GCGATCGC
 3098    GGCGATCG
 3046    CGATCGCC
 2554    CGATCGCG
 2494    CGCGATCG
 2328    AGCGATCG
 2316    CGATCGCT
 1860    CGATCGCA
 1845    TGCGATCG
 1071    GATCGCCC
 1042    GGGCGATC
  983    GATCGCGG
  960    GCGATCGG
  951    CCGATCGC
  910    GATCGCTG
  901    CCGCGATC
  900    TGGCGATC
  886    CAGCGATC
  876    AGGCGATC
  869    GATCGCCT
  863    GATCGCGA
  848    GATCGCAG
  844    GATCGCCA
  841    CTGCGATC
```

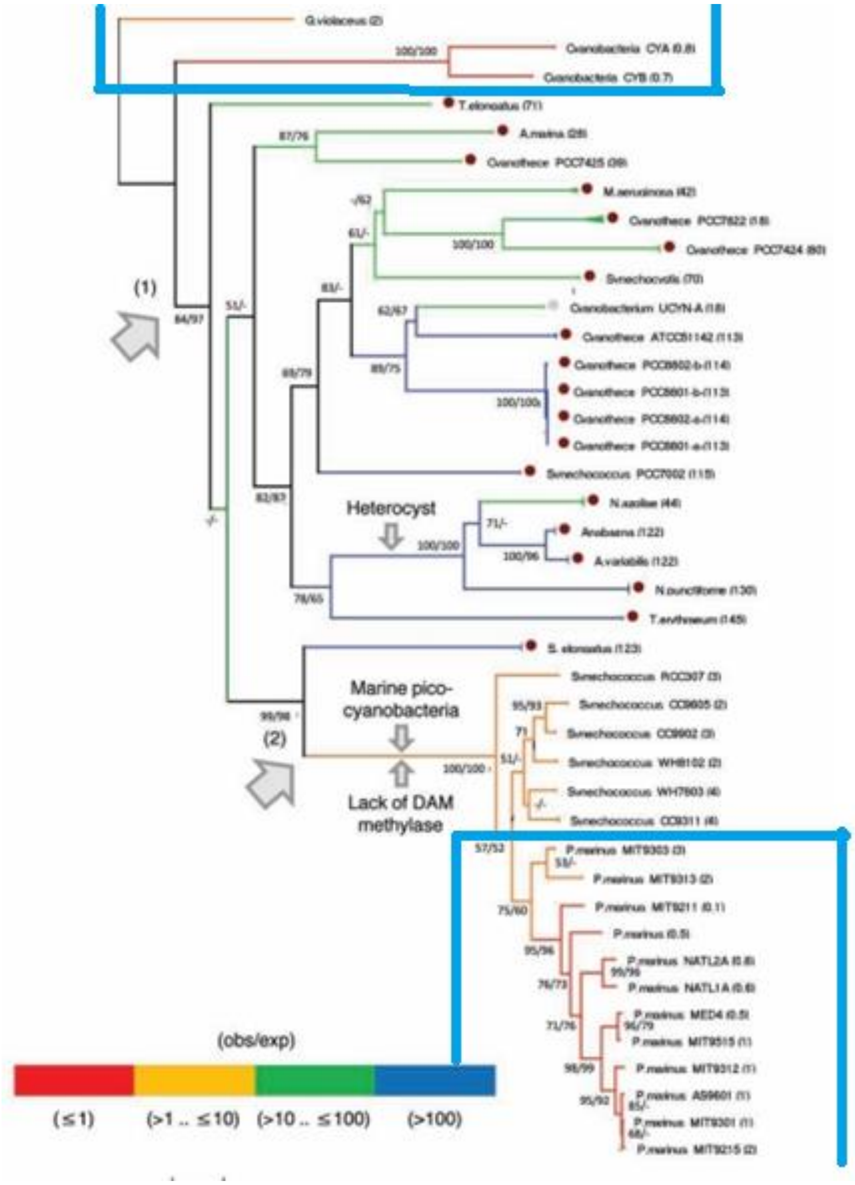
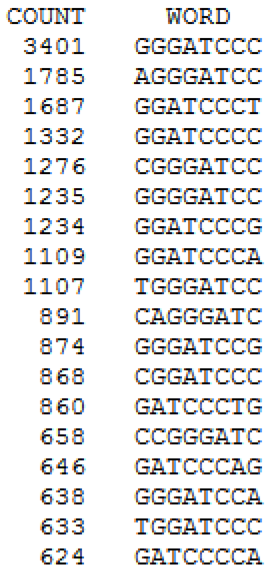
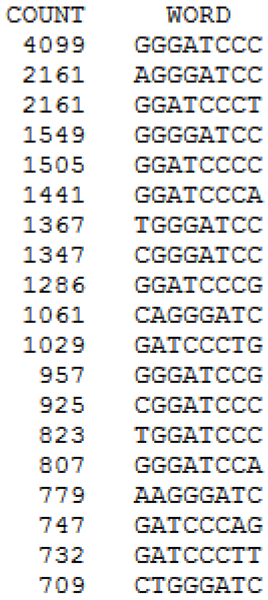


Figure 2. Phylogenic Tree (Delaye et al. 2011)



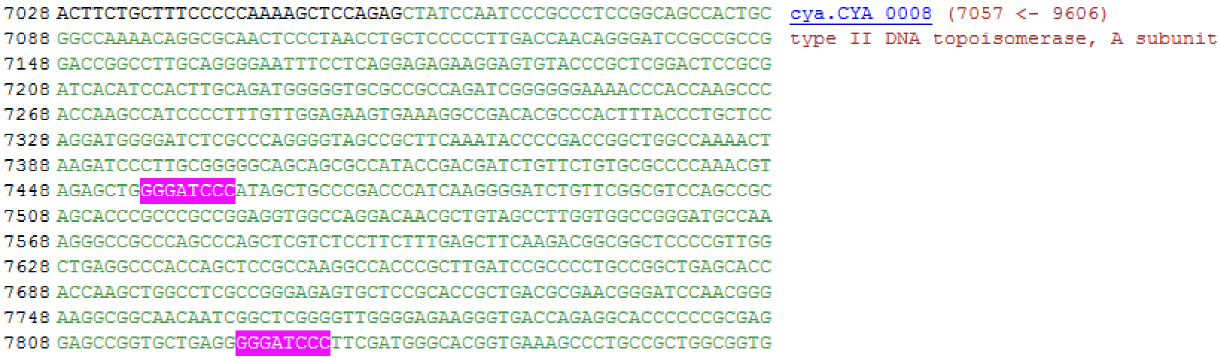
```
COUNT     WORD
 3401    GGGATCCC
 1785    AGGGATCC
 1687    GGATCCCT
 1332    GGATCCCC
 1276    CGGGATCC
 1235    GGGGATCC
 1234    GGATCCCG
 1109    GGATCCCA
 1107    TGGGATCC
  891    CAGGGATC
  874    GGGATCCG
  868    CGGATCCC
  860    GATCCCTG
  658    CCGGGATC
  646    GATCCCAG
  638    GGGATCCA
  633    TGGATCCC
  624    GATCCCCA
```

Figure 3: Most Common 8nt Sequences in CYA



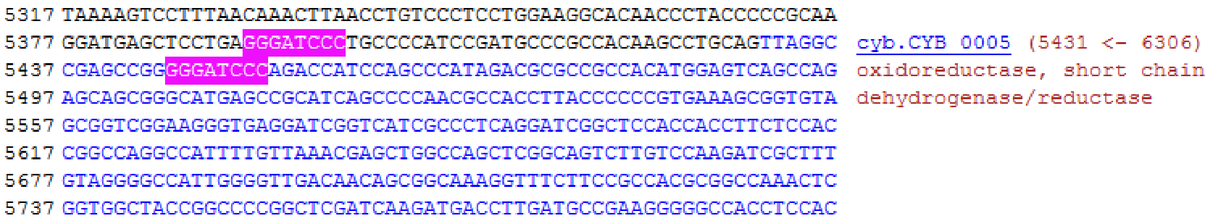
```
COUNT     WORD
 4099    GGGATCCC
 2161    AGGGATCC
 2161    GGATCCCT
 1549    GGGGATCC
 1505    GGATCCCC
 1441    GGATCCCA
 1367    TGGGATCC
 1347    CGGGATCC
 1286    GGATCCCG
 1061    CAGGGATC
 1029    GATCCCTG
  957    GGGATCCG
  925    CGGATCCC
  823    TGGATCCC
  807    GGGATCCA
  779    AAGGGATC
  747    GATCCCAG
  732    GATCCCTT
  709    CTGGGATC
```

Figure 4: Most Common 8nt Sequences in CYB



```
7028 ACTTCTGCTTTCCCCCAAAAGCTCCAGAGCTATCCAATCCCGCCCTCCGGCAGCCACTGC  cya.CYA_0008 (7057 <- 9606)
7088 GGCCAAAACAGGCGCAACTCCCTAACCTGCTCCCCCTTGACCAACAGGGATCCGCCGCCG  type II DNA topoisomerase, A subunit
7148 GACCGGCCTTGCAGGGGAATTTCCTCAGGAGAGAAGGAGTGTACCCGCTCGGACTCCGCG
7208 ATCACATCCACTTGCAGATGGGGGTGCGCCGCCAGATCGGGGGGAAAACCCACCAAGCCC
7268 ACCAAGCCATCCCCTTTGTTGGAGAAGTGAAAGGCCGACACGCCCACTTTACCCTGCTCC
7328 AGGATGGGGATCTCGCCCAGGGGTAGCCGCTTCAAATACCCCGACCGGCTGGCCAAAACT
7388 AAGATCCCTTGCGGGGGCAGCAGCGCCATACCGACGATCTGTTCTGTGCGCCCCAAACGT
7448 AGAGCTGGGGATCCCATAGCTGCCCGACCCATCAAGGGGATCTGTTCGGCGTCCAGCCGC
7508 AGCACCCGCCCGCCGGAGGTGGCCAGGACAACGCTGTAGCCTTGGTGGCCGGGATGCCAA
7568 AGGGCCGCCCAGCCCAGCTCGTCTCCTTCTTTGAGCTTCAAGACGGCGGCTCCCCGTTGG
7628 CTGAGGCCCACCAGCTCCGCCAAGGCCACCCGCTTGATCCGCCCCTGCCGGCTGAGCACC
7688 ACCAAGCTGGCCTCGCCGGGAGAGTGCTCCGCACCGCTGACGCGAACGGGATCCAACGGG
7748 AAGGCGGCAACAATCGGCTCGGGGTTGGGGAGAAGGGTGACCAGAGGCACCCCCCGCGAG
7808 GAGCCGGTGCTGAGGGGGATCCCTTCGATGGGCACGGTGAAAGCCCTGCCGCTGGCGGTG
```

Figure 5. Sample sequence of CYA gene containing found sequence.



```
5317 TAAAAGTCCTTTAACAAACTTAACCTGTCCCTCCTGGAAGGCACAACCCTACCCCCGCAA
5377 GGATGAGCTCCTGAGGGATCCCTGCCCCATCCGATGCCCGCCACAAGCCTGCAGTTAGGC  cyb.CYB_0005 (5431 <- 6306)
5437 CGAGCCGGGGGATCCCAGACCATCCAGCCCATAGACGCGCCGCCACATGGAGTCAGCCAG  oxidoreductase, short chain
5497 AGCAGCGGGCATGAGCCGCATCAGCCCCAACGCCACCTTACCCCCCGTGAAAGCGGTGTA  dehydrogenase/reductase
5557 GCGGTCGGAAGGGTGAGGATCGGTCATCGCCCTCAGGATCGGCTCCACCACCTTCTCCAC
5617 CGGCCAGGCCATTTTGTTAAACGAGCTGGCCAGCTCGGCAGTCTTGTCCAAGATCGCTTT
5677 GTAGGGGCCATTGGGGTTGACAACAGCGGCAAAGGTTTCTTCCGCCACGCGGCCAAACTC
5737 GGTGGCTACCGGCCCCGGCTCGATCAAGATGACCTTGATGCCGAAGGGGGCCACCTCCAC
```

Figure 6. Sample CYB sequence with found sequence in coding and non-coding region.

References

1) Delaye, L. Moya, A. Abundance and distribution of the highly iterated palindrome 1(HIP1) among prokaryotes Mob Genet Elements. 2011 Sep-Oct; 1(3) 159-168

2) Robinson, P. J., Cranenburgh, R. M., Head, I. M. and Robinson, N. J. (1997), HIP1 propagates in cyanobacterial DNA via nucleotide substitutions but promotes excision at similar frequencies in*Escherichia coli* and *Synechococcus* PCC 7942. Molecular Microbiology, 24: 181–189.

3) Robinson, P. J., Gupta, A., Bleasby, A., Whitton, B., Morby, AP. (1995) Singular over-representation of an octameric palindrome, HIP1, in DNA from many cyanobacteria. Nucleic Acids Research, 5: 729-735

4) Biobike Cyano http://biobike-8003.csbc.vcu.edu/biologin