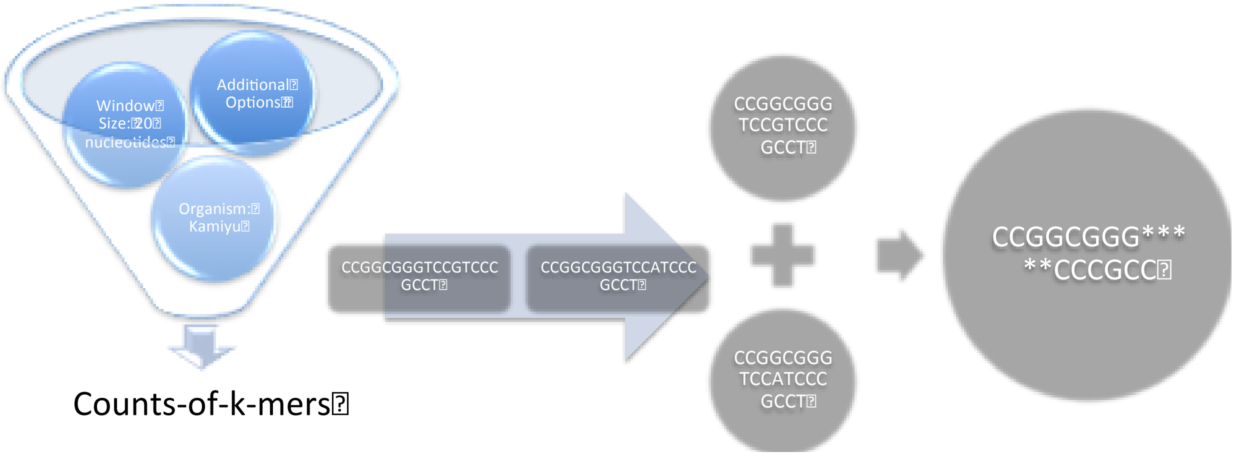
Introduction

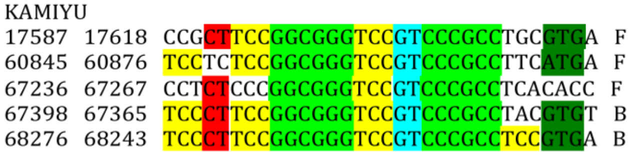
Let's start the expedition. Our group explored the realm of short dispersed repeats, and attempted to derive meaning from their functionality. Short dispersed repeats are approximately twenty to two hundred nucleotides long and occur in various genomes of both bacteria and viruses. For my portion of the project I decided to look at short dispersed repeats that occur in mycobacteriophages or viruses that infect mycobacteria such as Mycobacterium Tuberculosis and Mycobacterium Smegmatis which is a lab strain I have previously worked with. Since I was involved in Phages Discovered lab in the beginning of my sophomore year I have grown an interest in discovering what these viruses are capable of doing. Originally I was at a lose for where my starting point would be, and given my disdain for reading scientific papers I decided (rightly or wrongly) that I would go into the genome of mycobacteriophages and identify possible short dispersed repeats.

Methods



So initially I wanted to find something novel so my first choose was to look for twenty nucleotides repeats in mycobacteriophages using the counts-of-k-mers function in BioBike, which is visual interactive programming framework where students can analyze genomic sequences. The function counts-of-k-mers works by going through the genome of a specified organism and identifying repeats based on the sequence size the user designates. For example if I (the user) wanted to look for twenty nucleotides repeats the function will go through the genome and identify repetitive twenty nucleotide sequences, and generate a results pain that will display the actual repetitive sequences with the number of times the sequence repeats in the genome. In order test out whether this mechanism would be useful I picked a random mycobacteriophages and ran the counts-of-k-mer and defined it so that it would identify twenty nucleotide repetitive sequences. The mycobacteriophages I happened to choose was a sub-cluster B3 bacteriophage named Kamiyu. Using this function I found the short dispersed sequence "GGCGGGTCCGTCCCGCC", which occurred three times in mycobacteriopahge Kamiyu. I proceeded to use another function called matches-by-pattern, this function allows the user to input a sequence and an organism to look into for that sequence in the organism by matching its patterns in that target organism. I primarily used this function to get the coordinates of the sequence and identify locations of these sequences in Kamiyu. I was able to find additional sequences that were not identified in count-of-k-mers, and find interesting location specific tendencies. Although the sequence itself may not seem particularly remarkable the location of the repeats struck me as they mostly occurred at the tail end of the mycobacteriophage's sequence. I went through by hand and lined up the sequences on top of each other with the coordinates next to each one and looked for motifs, which I found as horrifically inefficient and there was actually a function that does that called motif in but the time had already passed so I moved on. When I did this I saw a hairpin sequences that occurred with a five-nucleotide long loops.

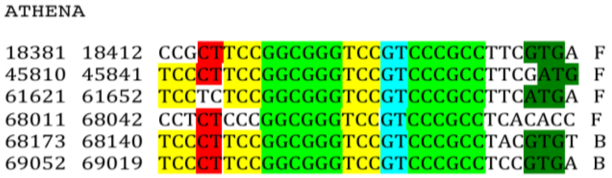
Key
Green: Palindrome (Suspect RNA Hairpin)
Yellow: Conserved short repeat of three nucleotides
Red: Partially Conserved duplicate
Blue: Partially Conserved duplicate



```
KAMIYU
17587  17618  CCGCTTCCGGCGGGTCCGTCCCGCCTGCGTGA  F
60845  60876  TCCTCTCCGGCGGGTCCGTCCCGCCTTCATGA  F
67236  67267  CCTCTCCCGGCGGGTCCGTCCCGCCTCACACC  F
67398  67365  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B
68276  68243  TCCCTTCCGGCGGGTCCGTCCCGCCTCCGTGA  B
```

For my next step I wanted to see if this sequence occurred in other bacteriophages in similar locations and in order to do this I used matches-by-pattern but looked at another phage in the same cluster as Kamiyu, called Athena and was able to see a similar pattern in location and sequence although insertions caused shifts further downstream.

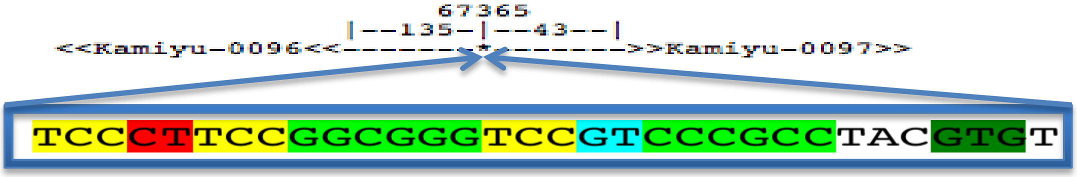
Key
Green: Palindrome (Suspect RNA Hairpin)
Yellow: Conserved short repeat of three nucleotides
Red: Partially Conserved duplicate
Blue: Partially Conserved duplicate



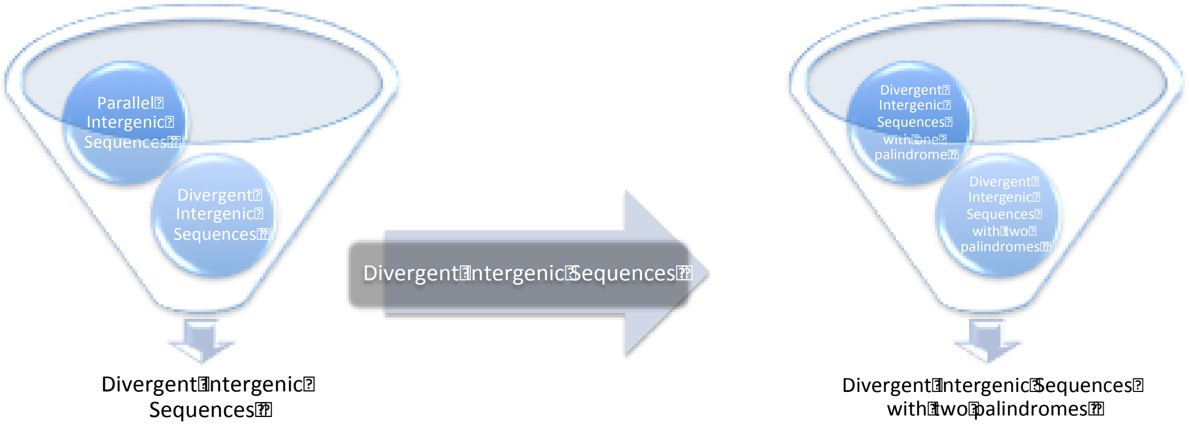
```
ATHENA

18381  18412  CCGCTTCCGGCGGGTCCGTCCCGCCTTCGTGA  F
45810  45841  TCCCTTCCGGCGGGTCCGTCCCGCCTTCGATG  F
61621  61652  TCCTCTCCGGCGGGTCCGTCCCGCCTTCATGA  F
68011  68042  CCTCTCCCGGCGGGTCCGTCCCGCCTCACACC  F
68173  68140  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B
69052  69019  TCCCTTCCGGCGGGTCCGTCCCGCCTCCGTGA  B
```

As to continue this process more efficiently I used the for-each loop function to map over the all mycobacteriophages list, which contained all the mycobacteriophage contained in the BioBike database. The for-each loop works by extracting each the name of each mycobacteriophages from the list of mycobacteriophages and inputting the phage into the body of the loop which was define as matches by pattern in this case and giving back similar results as above, but for each phage it actually found the sequence in on this the mycobacteriophages list. I only took phages where the sequence occurred in the same amount or more than Kamiyu, and in all cases those phages were in subcluster B3 phages that are available through biobike.


67365
|--135-|--43--|
<<Kamiyu-0096<<-------*------->>Kamiyu-0097>>
TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT

Actual nucleotide sequence is contained in the box above and the asterisk represents its context in terms of the first letter in the nucleotide sequence. The flanking regions are the genes it is in-between and the numbers to the left and right represent how far it is from the end or beginning of the reading frame for those particular proteins.

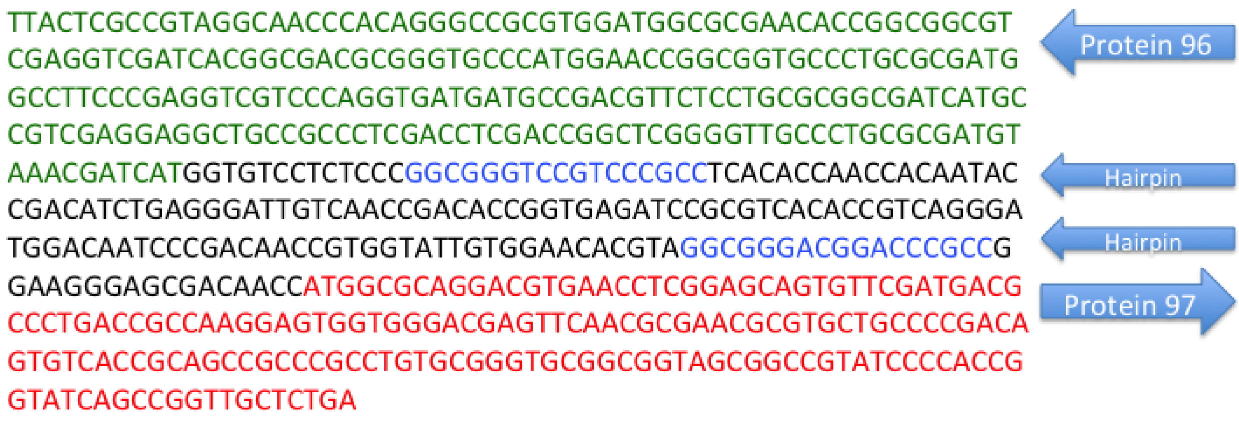

Parallel Intergenic Sequences
Divergent Intergenic Sequences
Divergent Intergenic Sequences
Divergent Intergenic Sequences
Divergent Intergenic Sequences with one palindrome
Divergent Intergenic Sequences with two palindromes
Divergent Intergenic Sequences with two palindromes

Filtered out only divergent intergenic sequences then re-filtered for divergent intergenic sequences with two palindromes

My next step was to look at the context of each of these sequences by taking the coordinates and their phage name and inserting it into the context-of function. This function works to tell the user where the sequence is in relation to the coding proteins of the genome. And, this occupied the bulk of my time but I ultimately did not user it thoroughly in my data generating given that it would return faulty matches that did not actually exist in the genomes. Although I still continued to use context of my extracting the sequence coordinates and mycobacteriophages names form the sequence similar to function. Many of the sequences were in genes that were transcribed parallel to on another, but some did occurring in genes that were divergent to one another and this is where my attention ended up shifting.

Results

I filter for only the divergent sequences and from there I filtered for the two divergent sequences that occurred in the same intergenic sequence for all the phages and in Kamiyu that happened to be protein 96 and protein 97 although that was not the case for the other phages, although for these two sequences the distance between any two protein there were present in was generally conserved.



```
TTACTCGCCGTAGGCAACCCACAGGGCCGCGTGGATGGCGCGAACACCGGCGGCGT
CGAGGTCGATCACGGCGACGCGGGTGCCCATGGAACCGGCGGTGCCCTGCGCGATG
GCCTTCCCGAGGTCGTCCCAGGTGATGATGCCGACGTTCTCCTGCGCGGCGATCATGC
CGTCGAGGAGGCTGCCGCCCTCGACCTCGACCGGCTCGGGGTTGCCCTGCGCGATGT
AAACGATCATGGTGTCCTCTCCCGGCGGGTCCGTCCCGCCTCACACCAACCACAATAC
CGACATCTGAGGGATTGTCAACCGACACCGGTGAGATCCGCGTCACACCGTCAGGGA
TGGACAATCCCGACAACCGTGGTATTGTGGAACACGTAGGCGGGACGGACCCGCCG
GAAGGGAGCGACAACCATGGCGCAGGACGTGAACCTCGGAGCAGTGTTCGATGACG
CCCTGACCGCCAAGGAGTGGTGGGACGAGTTCAACGCGAACGCGTGCTGCCCCGACA
GTGTCACCGCAGCCGCCCGCCTGTGCGGGTGCGGCGGTAGCGGCCGTATCCCCACCG
GTATCAGCCGGTTGCTCTGA
```

Region with blue lettering represents the dual palindromic sequences in the intergenic region between Kamiyu Protein 96 and Protein 97

Interestingly enough there were not convergent sequences where their palindromes were located which may be a indictor to what their functions will be.

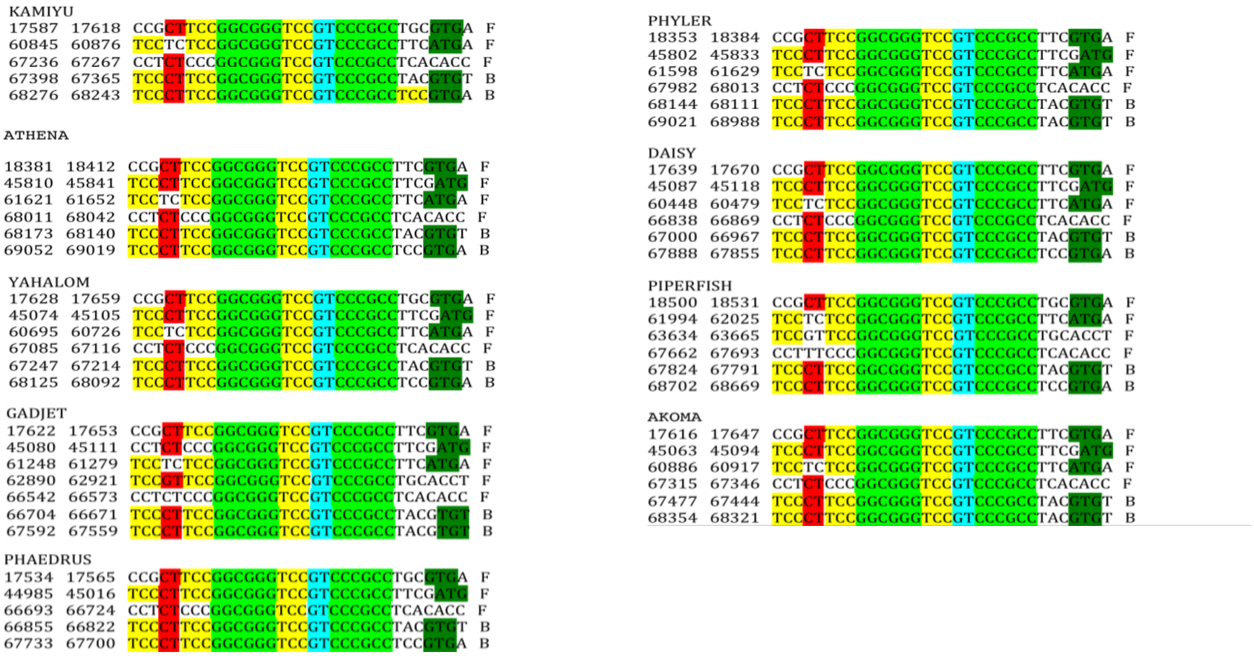
Discussion

Possibly that most difficult portion of this project was determining what the purpose of my actual sequences were. Just looking at the region with their palindromes I looked at the domain of the upstream and downstream proteins to see if there were any prediction functions and I was able to find a coiled-coil structure upstream in Kamiyu and a membrane protein structure downstream from the sequence in Kamiyu. Another thing that I did not have time to do additional research on but observed recently was that although the identified sequences occur in genes transcribed parallel to the sequence there was an interesting overrepresentation of that sequence in divergent regions

My view of these sequences is that they may serve as genetic switches similar to cro and cI of lambda phage or directional switches that control protein production that may be related to their lytic of lysogenic phase of the phage.

Appendix

Motifs in Original Phage Cluster



```
KAMIYU
17587  17618  CCGCTTCCGGCGGGTCCGTCCCGCCTGCGTGA  F
60845  60876  TCCTCTCCGGCGGGTCCGTCCCGCCTTCATGA  F
67236  67267  CCTCTCCCGGCGGGTCCGTCCCGCCTCACACC  F
67398  67365  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B
68276  68243  TCCCTTCCGGCGGGTCCGTCCCGCCTCCGTGA  B

ATHENA

18381  18412  CCGCTTCCGGCGGGTCCGTCCCGCCTTCGTGA  F
45810  45841  TCCCTTCCGGCGGGTCCGTCCCGCCTTCGATG  F
61621  61652  TCCTCTCCGGCGGGTCCGTCCCGCCTTCATGA  F
68011  68042  CCTCTCCCGGCGGGTCCGTCCCGCCTCACACC  F
68173  68140  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B
69052  69019  TCCCTTCCGGCGGGTCCGTCCCGCCTCCGTGA  B

YAHALOM
17628  17659  CCGCTTCCGGCGGGTCCGTCCCGCCTGCGTGA  F
45074  45105  TCCCTTCCGGCGGGTCCGTCCCGCCTTCGATG  F
60695  60726  TCCTCTCCGGCGGGTCCGTCCCGCCTTCATGA  F
67085  67116  CCTCTCCCGGCGGGTCCGTCCCGCCTCACACC  F
67247  67214  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B
68125  68092  TCCCTTCCGGCGGGTCCGTCCCGCCTCCGTGA  B

GADJET
17622  17653  CCGCTTCCGGCGGGTCCGTCCCGCCTTCGTGA  F
45080  45111  CCTCTCCCGGCGGGTCCGTCCCGCCTTCGATG  F
61248  61279  TCCTCTCCGGCGGGTCCGTCCCGCCTTCATGA  F
62890  62921  TCCGTTCCGGCGGGTCCGTCCCGCCTGCACCT  F
66542  66573  CCTCTCCCGGCGGGTCCGTCCCGCCTCACACC  F
66704  66671  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B
67592  67559  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B

PHAEDRUS
17534  17565  CCGCTTCCGGCGGGTCCGTCCCGCCTGCGTGA  F
44985  45016  TCCCTTCCGGCGGGTCCGTCCCGCCTTCGATG  F
66693  66724  CCTCTCCCGGCGGGTCCGTCCCGCCTCACACC  F
66855  66822  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B
67733  67700  TCCCTTCCGGCGGGTCCGTCCCGCCTCCGTGA  B
```

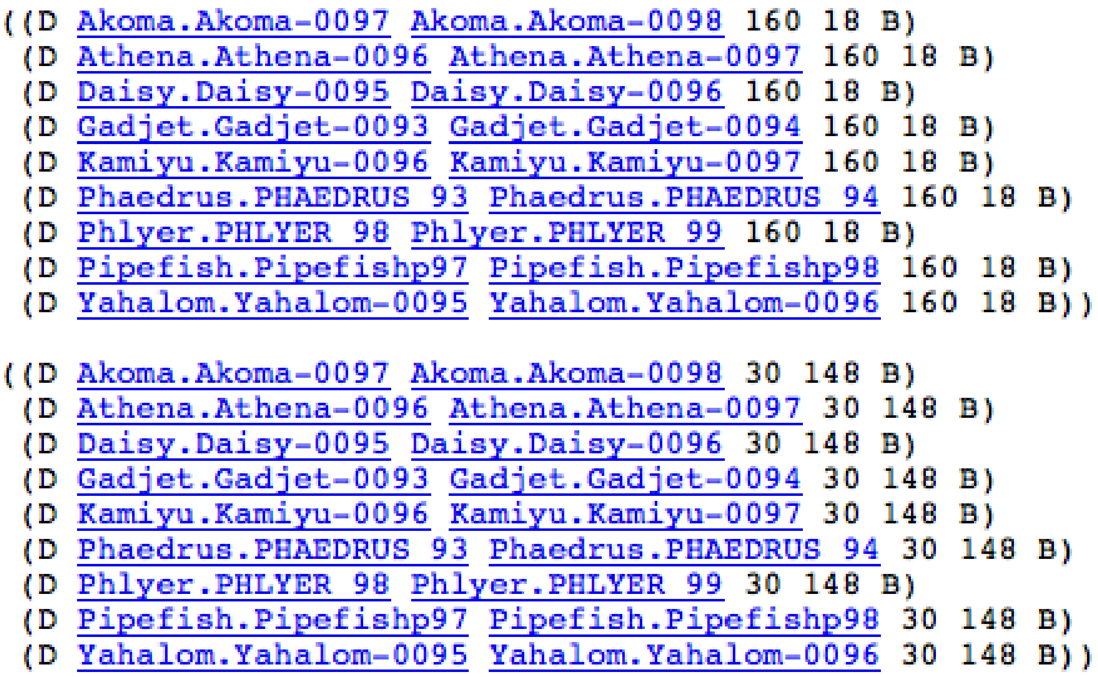
```
PHYLER
18353  18384  CCGCTTCCGGCGGGTCCGTCCCGCCTTCGTGA  F
45802  45833  TCCCTTCCGGCGGGTCCGTCCCGCCTTCGATG  F
61598  61629  TCCTCTCCGGCGGGTCCGTCCCGCCTTCATGA  F
67982  68013  CCTCTCCCGGCGGGTCCGTCCCGCCTCACACC  F
68144  68111  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B
69021  68988  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B

DAISY
17639  17670  CCGCTTCCGGCGGGTCCGTCCCGCCTTCGTGA  F
45087  45118  TCCCTTCCGGCGGGTCCGTCCCGCCTTCGATG  F
60448  60479  TCCTCTCCGGCGGGTCCGTCCCGCCTTCATGA  F
66838  66869  CCTCTCCCGGCGGGTCCGTCCCGCCTCACACC  F
67000  66967  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B
67888  67855  TCCCTTCCGGCGGGTCCGTCCCGCCTCCGTGA  B

PIPERFISH
18500  18531  CCGCTTCCGGCGGGTCCGTCCCGCCTGCGTGA  F
61994  62025  TCCTCTCCGGCGGGTCCGTCCCGCCTTCATGA  F
63634  63665  TCCGTTCCGGCGGGTCCGTCCCGCCTGCACCT  F
67662  67693  CCTTTCCCGGCGGGTCCGTCCCGCCTCACACC  F
67824  67791  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B
68702  68669  TCCCTTCCGGCGGGTCCGTCCCGCCTCCGTGA  B

AKOMA
17616  17647  CCGCTTCCGGCGGGTCCGTCCCGCCTTCGTGA  F
45063  45094  TCCCTTCCGGCGGGTCCGTCCCGCCTTCGATG  F
60886  60917  TCCTCTCCGGCGGGTCCGTCCCGCCTTCATGA  F
67315  67346  CCTCTCCCGGCGGGTCCGTCCCGCCTCACACC  F
67477  67444  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B
68354  68321  TCCCTTCCGGCGGGTCCGTCCCGCCTACGTGT  B
```

Dual Palindrome Sites



```
((D Akoma.Akoma-0097 Akoma.Akoma-0098 160 18 B)
 (D Athena.Athena-0096 Athena.Athena-0097 160 18 B)
 (D Daisy.Daisy-0095 Daisy.Daisy-0096 160 18 B)
 (D Gadjet.Gadjet-0093 Gadjet.Gadjet-0094 160 18 B)
 (D Kamiyu.Kamiyu-0096 Kamiyu.Kamiyu-0097 160 18 B)
 (D Phaedrus.PHAEDRUS 93 Phaedrus.PHAEDRUS 94 160 18 B)
 (D Phlyer.PHLYER 98 Phlyer.PHLYER 99 160 18 B)
 (D Pipefish.Pipefishp97 Pipefish.Pipefishp98 160 18 B)
 (D Yahalom.Yahalom-0095 Yahalom.Yahalom-0096 160 18 B))

((D Akoma.Akoma-0097 Akoma.Akoma-0098 30 148 B)
 (D Athena.Athena-0096 Athena.Athena-0097 30 148 B)
 (D Daisy.Daisy-0095 Daisy.Daisy-0096 30 148 B)
 (D Gadjet.Gadjet-0093 Gadjet.Gadjet-0094 30 148 B)
 (D Kamiyu.Kamiyu-0096 Kamiyu.Kamiyu-0097 30 148 B)
 (D Phaedrus.PHAEDRUS 93 Phaedrus.PHAEDRUS 94 30 148 B)
 (D Phlyer.PHLYER 98 Phlyer.PHLYER 99 30 148 B)
 (D Pipefish.Pipefishp97 Pipefish.Pipefishp98 30 148 B)
 (D Yahalom.Yahalom-0095 Yahalom.Yahalom-0096 30 148 B))
```