Introduction

Bacteria as a domain are highly resilient organisms. Struggles for food, resources and hospitable living conditions are a constant battle for bacteria, but they manage to proliferate at extreme rates despite these limiting conditions. However, one particular threat can be highly damaging to a bacterial population, even in an otherwise friendly environment. That threat is the phage. Specifically a bacterial virus, phages have incredible population size, with an estimated $10^{30}$ total phages in the biosphere[1] (that is one million trillion trillion). In fact, samples of sea waters show up to 70% of observed bacteria to be infected with phages, with 900 million virons existing per milliliter of sea water[2].

In order to deal with this constant and ubiquitous threat, bacteria have evolved a number of ways to combat phages. One particularly interesting strategy bacteria employ is one that is a homoplasic to the B cell-mediated, antibody-based adaptive immune system found in mammals. However the adaptive immune system of the bacteria is based immediately off of the bacteria's own genome. This immune response is called a CRISPR, or a "clustered regularly interspaced short palindromic repeat". These are regions in the DNA which Have repeating sequences 23-47 base pairs long, interrupted by variable regions of similar or slightly longer legnth[3]. These two regions which sandwich each other are referred to as direct repeat sequences, and variable spacer sequences.

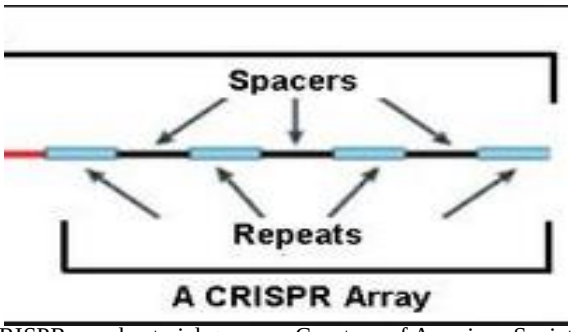



Figure 1. Diagram of a basic CRISPR on a bacterial genome. Courtesy of American Society for Microbiology website, http://schaechter.asmblog.org/schaechter/2011/04/six-questions-about-crisprs.html

CRISPRS function by transcribing a spacer sequence into RNA and associating it with a Cas (CRISPR-associated protein. Via this method, the CRISPR can effectively act as a search tool to find DNA sequences like itself[4]. What it is searching for is phage DNA. If the CRISPR finds a match, it can tag the phage DNA for deletion, and potentially save the bacteria from infection.

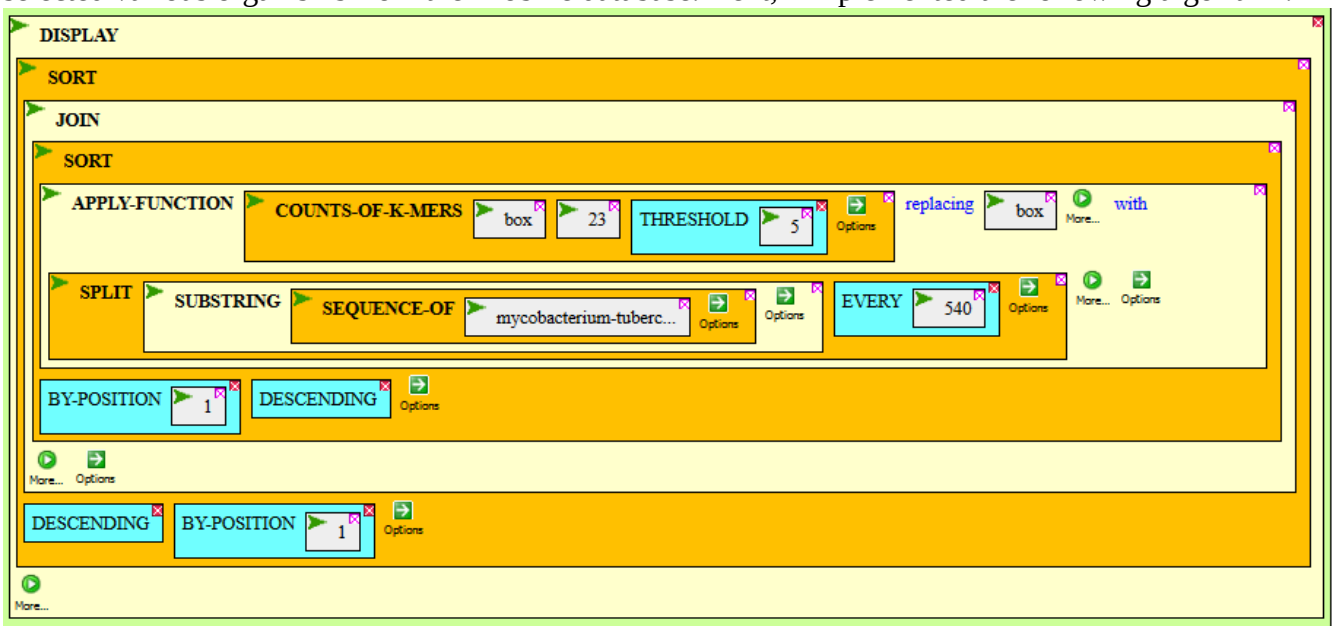
However, the bacteria needs to know what to search for first. The variable spacer sequences come directly from free DNA the bacteria has previously encountered, and assimilated into its genome. While this DNA may be from non-phage sources, it is often phage DNA[5]. By this method, if a bacteria has encountered phage DNA and not been killed, it may gain active immunity towards that phage, as well as all of its exponentially reproducing descendants.

Due to this consideration, there should be large similarities between bacteria CRISPR variable spacer sequences and actual phage DNA.

To look into this, it was my goal to first find CRISPR sequences. My next goal was to, after finding such sequences, compare the variable spacer sequences to known phage DNA. My question was twofold: If I could find a CRISPR variable sequence that matches to the genome of a reasobale phage candidate, and second: If the collection of variable regions in CRISPRs has any bias towards certain characteristics of a virus genome 9ie; coding vs noncoding regions).

Methods

As previously mentioned, the first step is to actually find a CRISPR sequence. To do this, I randomly selected various organisms from the Biobike database. Next, I implemented the following algorithm:

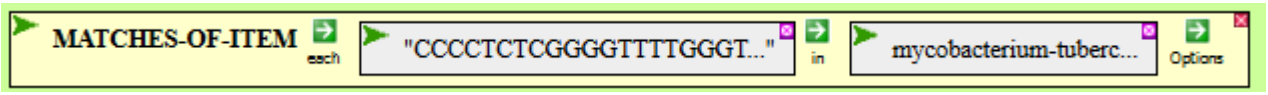



This breaks the bacterium's genome into chunks of 540. I chose this box size for two reasons. 1, previous work on a completely unrelated project happened to come across a CRISPR sequence by happenstance, and a box size of 540 was being used then (with slightly better justification as well). The second reason is that 540 is actually a fairly good search size. CRISPR sequences appear in tight clusters in the genome, not spread throughout the chromosome. Assuming a maximum sized CRISPR, with a direct repeat sequence of 47bps and a variable spacer sequence of 60 bps, I should still return at least five results.

The program then, in each 540 chunk, executes the “counts-of-k-mers” function. This tallies up nucleotide strings of 23bps in length that repeat multiple times. Here, I set 23 as the search size as this is the lower limit for CRISPR direct repeat regions. I also set the threshold to 5, as this would return more likely results (as previously stated, 5 should be the minimum return even for a large CRISPR). The compiled list was then joined and sorted to be more easily viewable. In this example, using Mycobacterium-tuberculosis-H37Ra, a maximum result of 7 was obtained.

| COUNT | WORD |
|---|---|
| 7 | CCCCTCTCGGGGTTTTGGGTCTG |
| 7 | CCCTCTCGGGGTTTTGGGTCTGA |
| 7 | CCGTCCCCTCTCGGGGTTTTGGG |
| 7 | CCTCTCGGGGTTTTGGGTCTGAC |
| 7 | CGTCCCCTCTCGGGGTTTTGGGT |
| 7 | CTCGGGGTTTTGGGTCTGACGAC |
| 7 | CTCTCGGGGTTTTGGGTCTGACG |
| 7 | GTCCCCTCTCGGGGTTTTGGGTC |
| 7 | GTTTCCGTCCCCTCTCGGGGTTT |
| 7 | TCCCCTCTCGGGGTTTTGGGTCT |
| 7 | TCCGTCCCCTCTCGGGGTTTTGG |
| 7 | TCTCGGGGTTTTGGGTCTGACGA |
| 7 | TTCCGTCCCCTCTCGGGGTTTTG |
| 7 | TTTCCGTCCCCTCTCGGGGTTTT |
| COUNT | WORD |
| 6 | CCCCTCTCGGGGTTTTGGGTCTG |
| 6 | CCCTCTCGGGGTTTTGGGTCTGA |
| 6 | CCGTCCCCTCTCGGGGTTTTGGG |
| 6 | CCTCTCGGGGTTTTGGGTCTGAC |
| 6 | CGTCCCCTCTCGGGGTTTTGGGT |

From here, the sequence can be selected and searched for in the bacterium's genome using matches-of-item.




The specific instance above (highlighted) contained a highly probable CRISPR. The following is the returned sequence from the bacterium. It is also interrupted by an unknown protein. This is interesting, but has little to do with the topic at hand.

ATGCTCATCCTGAATGCCGGTCAACAGACGCGGTGGCGACCCAGTCGTCGTAGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTG
ACGACTCGGGCACGGCCGAAACACCGCGCGAAGGGCGGTTCAAGTTTCCGTCCCCTCTCGTGGTTTTGGGTCTGACGACTGGG
AGGATGTCACTCGGACATAGCTGTCATCGGCGGTGTGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGAGGACATGGAGCAGTA
GCGTGGCTGTGGTGTGGCGGGCGATATGCGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACTGCTGCACCTCCCGCACC
CGGTGCGATTCTGCGTCCAGTTTCCGTCCCCTCTCGGGGTTTTGGGTCCGACGACCCCGATAGTCGCGCTCGTCCATGTCCCA
CCATGAGGGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACTACCTGATAGAAGCCGGAAAGCTCCGTGCCGTCAGGTTT
CCGTCCCCTCTCGGGGTTTTGGGTCTGACGACAGGGCACTGGACCTGTATGAGGCACAGATGGCGTACTAGTTTCCGTCCCCT
CTCGGGGTTTTGGGTCTGACGACCCGGATCGGTTACCCACGCCGATTTACTGGCCATCGTCGGGTTTCCGTCCCCTCTCGGGG
TTTTGGGTCTGACGACACTTGCGCGCACAACGCATCCGCCATCCACGGGGCGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGA
CGACCTGAAAGGGGGACTGTGGACGAGTTCGCGCTCAAAATGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACTTGAAC
ACGCCGATACCTATTTGGTCGGGAGTGATAAAGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACCGGACTTGATCGACG
CGAACCTGTCTGACGCGAACCTGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACGGCTGGAAAAGGGCGCGGGGCAACC
GCATCGTCAAGAGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACGCGTTGTGGTCGTGTCGTGGAGCCTGTATTTCGCT
GGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACCATTAGTTGGTGTTGTGATCGCTAAACGCCGGGGCAGTTTCCGTCC
CCTCTCGGGGTTTTGGGTCTGACGACCTATCCGCGGGAAGAGATCACGAATCCGGCGTCGAAGGGTTTCCGTCCCCTCTCGGG
GTTTTGGGTCTGACGACATGCTGAGCTGAGGCGCCGGATGATGGTGGTGCTGAAGGTTTCCGTCCCCTCTCGGGGTTTTGGGT
CTGACGACTGACAGGGTGCGGTGGTCGCTGATCGGCTCCCCGAGTTTCCGTCCCCTCTCGGGGTGAACCGCCCCGGTGAGTCC
GGAGACTCTCTGATCTGAGACCTCAGCCGGCGGCTGGTCTCTGGCGTTGAGCGTAGTAGGCAGCCTCGAGTTCGACCGGCGGG
ACGTCGCCGCAGTACTGGTAGAGGCGGCGATGGTTGAACCAGTCGACCCAGCGCGCGGTGGCCAACTCGACATCCTCGATGGA
CCGCCAGGGCTTGCCGGGTTTGATCAGCTCGGTCTTGTATAGGCCGTTGATCGTCTCGGCTAGTGCATTGTCATAGGAGCTTC
CGACCGCTCCGACCGACGGTTGGATGCCTGCCTCGGCGAGCCGCTCGCTGAACCGGATCGATGTGTACTGAGATCCCCTATCC
GTATGGTGGATAACGTCTTTCAGGTCGAGTACGCCTTCTTGTTGGCGGGTCCAGATGGCTTGCTCGATCGCGTCGAGGACCAT
GGAGGTGGCCATCGTGGAAGCGACCCGCCAGCCCAGGATCCTGCGAGCGTAGGCGTCGGTGACAAAGGCCACGTAGGCGAACC
CTGCCCAGGTCGACACATAGGTGAGGTCTGCTACCCACAGCCGGTTAGGTGCTGGTGGTCCGAAGCGGCGCTGGACGAGATCG
GCGGGACGGGCTGTGGCCGGATCAGCGATCGTGGTCCTGCGGGCTTTGCCGCGGGTGGTCCCGGACAGGCCGAGTTTGGTCAT
CAGCCGTTCGACGGTGCATCTGGCCACCTCGATGCCCTCACGGTTCAGGGTTAGCCACACTTTGCGGGCACCGTAAACACCGT
AGTTGGCGGCGTGGACGCGGCTGATGTGCTCCTTGAGTTCGCCATCGCGCAGCTCGCGGCGGCTGGGCTCCCGGTTGATGTGG
TCGTAGTAGGTCGATGGGGCGATCGGCACACCCAGCTCGGTCAGCTGTGTGCAGATCGACTCGACACCCCACCGCAAACCATC
GGGGCCCTCGCGGTGGCCCTGATGATCGGCGATGAACCGGGTAATTAGCGTGCTGGCCGGTCGAGCTCGGCCGCGAAGAAAGC
CGACGCGGTCTTTAAAATCGCGTTCGCCCTTCGCAATTCGGCGTTGTCCCGCCGCAAGCGCTTCAGCTCAGCGGATTCTTCGG

TCGTGGTCCCGGGCCGTGCGCCGGCATCGACCTGCGCCTGGCGCACCCACTTACGCACCGTCTCCGCGCAGCCAACACCAAGT
AGACGGGCGACCTCACTGATCGCTGCCCACTCCGAATCGTGCTGACCGCGGATCTCTGCGACCATCCGCACCGCCCGCTCACG
CAGCTCCGGCGGGTACCTCCTCGATGAACCACCTGACATGACCCCATCCTTTCCAAGAACTGGAGTCTCCGGACATGCCGGGG
CGGTTCAGGGTTTTGGGTCTGACGACTCGCGGCGAGCACGTCTCACCCAGCAGGCGGTGAGGTTGGGTTTCCGTCCCCTCTCG
GGGTTTTGGGTCTGACGACACGGACGAGCTGGACCGCATCAGCGATGCTGAGCTGAGGGTTTCCGTCCCCTCTCGGGGTTTTG
GGTCTGACGACTTGTCTCAATCGTGCCGTCTGCGGTGACACGCTCCAAGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGA
CCACCAGGATCAGCGCCAAGCCAGTTAGCGCAATCCAGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACCTCCCGGACC
ATCTGCAGCTCGCCCGGGTCCATGCGGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACCGGAGTCATCCGCGCGGGCCG
GCGCGATTGTTGCCGGGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACTGGCGATTTACGACGCTGACGGGAACTCGTG
CGAATGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGATCCGCGAAATTCACTGCGCGTTATTCAAGGTTTCCGTCCCCTCTCG
GGGTTTTGGGTCTGACGACCCGAGCCGACCATCCGCATCACACCGAAAGGGTTGGCGCAAGTTTCCGTCCCCTCTCGGGGTTT
TGGGTCTGACGACACGTGGGGAGAGGGAATGGCAATGATGGTCGACGAAGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACG
ACCTCGGACAGCATCTCCCCGGGCGGGCAGCAGATATCCCATGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACCGACC
CGTGGCCGCCAGGTTGCCGCCGCCGTTGCTCACCTGGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACCCGGAAGTCAA
CTAGAGCGGGTGTCGAACGCTGCCCGGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACATGCGAATCCGCTGTCAGCAC
ATGGGATTCCGAGTGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACCTAGGCGGCCCCGGCGAGGCTGGGGGCGGTTTC
ACGCGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACCAGCGCAGACGGCAGCCCCGAGTACTCGCTCTCCTCAGGTTTC
CGTCCCCTCTCGGGGTTTTGGGTCTGACGACAGGCTGAAATTGAAGCCGGAAATGACGACGCATTGGTGTTTCCGTCCCCTCT
CGGGGTTTTTGGGGTCTTGACGACCTAAAGCCCCGCCTAAATCCCCGCACAAAGTTGGGTCAGAAAAAAGGTTTTCCCGTTCC
CCCCTCCTCGGGGGGGTTTTGGGTCTGACGACCTGATGATTGGTCGGCGTATGACGTGCTACTGAGGTGTTGTTTCCGTCCCC
TCTCGGGGTTTTGGGTCTGACGACTAGAAGGCGATCACTGGAAGCACGGCGCTTGCGAGTTTCCGTCCCCTCTCGGGGTTTTG
GGTCTGACGACTTGGTCAAAAGCTGTCGCCCAAGCATGAGGCAAAAAGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGAC
ACGACTAGGGGAGCGTGATCCAGAGCCGGCGACCCTCTATGGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACGTGCAA
GAATTCCGGGTTGCAGTGCAACACGGTTTTAAGTTTCCGTCCCCTCTCGGGGTTTTGGGTCTGACGACTCTATGGACAATTCG
TCCAGCGTGTGGTAACAATGCCTGCTGATGATGTCAAAAGAACACAAACTCCTCTGCGCTGACAAGCCGTCCCCTTCCGTAGA
ACGTAACTGCCGCAACACCTCTTATCTTATAGATCCGGATGTTGTCGCAGTCGATGGCGAAGCGGTCGATACGTGCAACTAGT
TTCGCGAGCTGGCCCTTCGTCAGCATCGCTTCGAATGCGGACTC

As can be seen, the direct repeat sequence is 36bps in length. Based on Biobike's knowledge of protein function, the sequence is also directly before a CRISPR protein.

```
3135481 TGGGTCTGACGACGTGCAAGAATTCCGGGTTGCAGTGCAACACGGTTTTAAGTTTCCGTC
3135541 CCCTCTCGGGGTTTTGGGTCTGACGACTCTATGGACAATTCGTCCAGCGTGTGGTAACAA
3135601 TGCCTGCTGATGATGTCAAAAGAACACAAACTCCTCTGCGCTGACAAGCCGTCCCCTTCC  Mtub-H37Ra.Mtub-H37Ra-7901 (3135616 <- 3135957)
3135661 GTAGAACGTAACTGCCGCAACACCTCTTATCTTATAGATCCGGATGTTGTCGCAGTCGAT  CRISPR-associated protein Cas2
3135721 GGCGAAGCGGTCGATACGTGCAACTAGTTTCGCGAGCTGGCCCTTCGTCAGCATCGCTTC
3135781 GAATGCGGACTCTTGGACGCGATAGCCAAACCCGGCCAGGATCTTCGCAAGTGAAGCCCG
3135841 CCGCCGGTTGTCGCTGATGTCGTATATTACGAGGACGAACATCTTGCCTATAGTGCCGCT
3135901 GGACTCGTCCACTTTGAGCGGGAGATTGAAGTACTCCTCACGGCTGCGAGTGGGCATTTA  Mtub-H37Ra.Mtub-H37Ra-7902 (3135958 <- 3136974)
3135961 GGCTCCGGATGGCTCGGAGGTGATATCGATATCGACGAGCCGCGACGGGTGCCCGGCTTC  CRISPR-associated protein Cas1 [SS]
3136021 GATAACACGCACGAGGCTTTGCAGTTGCAAGTCGAGGGCGTACTGAAAGGTGTATCGGTG
3136081 AGGATCGCCTTTGATGTAGGTGGCGGTTCGTGCGATTCGATTACCAAAGGCGCGCGCGAT
3136141 GGATCGTGTGGCTTCCCGTGTCGCGAAGACGGCCCCCGTGTCGGAGTTCTTGCTGAAAGC
3136201 CCGGGTGTCGACCACACCGTCCGCGATCAATCGAAGTACGGTGTCATCGATGATCGGCGC
3136261 CCGCCATACCTCCATGAGGTCGCTCGCCAACGTTGCGTGCCCTCGTGAATCCTGGTGTAG
3136321 GAAACCGATATACGCGTTCAGGCTGTGACGCTCGATCGCCCCTATGATGTTCTTGTACAG
3136381 CAGCGAATAGCCGAGGCTGACCATCGAGTTGAAGGCGTCCAACGGCGGCCGAGTCGAGCG
3136441 GCCCTGGAATGCGAACTCCTGCGGGACGAGATGCCCCAGCGCGGTGAAGTATGCCTTTGC
3136501 GGCATTTCCCTCGAACCCGTTCAACTCCGCCAGGGAGCCCGATCGATCGACCCAGGCCAG
3136561 CGAGTGCTTCATCGTGCGGATGCTCTCAGCAACGTCTTGCCCCGACGTGTGTGCCCGAAT
3136621 CAAGGCCTGCTGATTCAGGATCTTCCTCGACACGATCCGCTTGCTTAACGACAGGCAGAA
3136681 CGCAGGATCGTCGGTGCGGTGAACTTGCTGACGGAGCCGCGGCGCGTATGACACGTCGGG
3136741 TGTTGAGATCCGGCCCTGGTAGTGGCCGTCGGTCGTGAAGAGCTGGATGTCGCGCTCACG
3136801 CTTGAGCATCTCAACGATGAAGGGCGTTGTCATCGTCGGCCGCCCAAACAGCGTGATGCC
```

So, this region is very likely to be a CRISPR. The method of finding CRISPRs appears to be effective.

I repeated this with other bacterial organisms.

Nitrosomonas europaea, a nitrogen fixing bacterium found in soils, returned this:

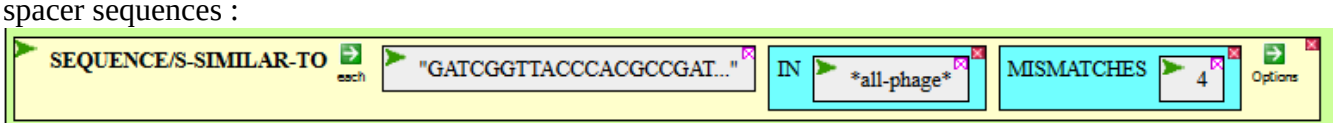
```
AACGCATTTACGCACTAATCCGGCGGTAATTCAGGGTATCGGTGTTTCGGTATCCTGCTGTTGAAAAAG
ATAGGTGGGCGTTGTCTCAATCCCTTTGAAATCAGGGCATCGGTGTTTCATCTATTGGCCGTATGGTTT
CGAAGTAAAGATGGTCTCAATCCCTTTGAAATCAGGGCATCGGTGTTTCTCTACTGGTTCGAGACATAT
GAGGAAAAGTATTTGTCTCAATCCCTTTGAAATCAGGGCATCGGTGTTTCAAACCAGAATCAAGAAAGA
AGACAAAAATTACTACGTCTCAATCCCTTTGAAATCAGGGCATCGGTGTTTCAAACCAGAATCAAGAAA
GAAGACAAAAATTACTACGTCTCAATCCCTTTGAAATCAGGGCATCGGTGTTTCCCTTGGCTAACGCGA
```

And C. tetani, which causes tetnus, returned this:

GGGAGTAGACAATTATAGTTATATTAATGGTTTATAAATTATTTAACATAACTTGTATTATAATA
ATAGACTATAAGAAATCCCAACTTAGTAGGTAGACAGAAATAGAAGTTATACTATAGACATAT
AGTGCATCATAGATGAACGCGTATTAGTAGCACCATATTGGAATGTAAATTTAAATGAAGGTA
CTAAATTTAAGGTAAGAATGGTGGTATTAGTAGCACCATATTGGAATGTAAATAGCATTCCTC
TATCTCCATTAACTACTGAAAAAGGAGTATTAGTAGCACCATATTGGAATGTAAATGCTATAA
GAATAATTCTAATTTTATCCAAGGAACTGGTATTAGTAGCACCATATTGGAATGTAAATTTCC
AATTTTACTAGCTGCTACCCCAACGCCTAATAATGTATTAGTAGCACCATATTGGAATGTAAA
TTAAATATATTACTTCCTTCTTGCACTGTAGGTTTTTCGTATTAGTAGCACCATATTGGAATGT
AAATTTGTTCGTAACTGTTAAAGCTATCTTTCTTTATGCGTATTAGTAGCACCATATTGGAATG
TAAATAGAGCCTAGTTTCCTAAGCCCTTATAACCAACTTACCGTATTAGTAGCACCATATTGG
AATGTAAATATACAATGCTCCATGGAAAGGACTCCACTTAGATATATAGTATTAGTAGCACCA
TATTGGAATGTAAATCCCACATCATTAAAGGATATAAAATTACCACCTTCCGTATTAGTAGCA
CCATATTGGAATGTAAATGTTTTAATATTAATATCGGCAAGTGCTAATTCATATGGTATTAGTA
GCACCATATTGGAATGTAAATGTTTTAATATTAATATCGGCAAGTGCTAATTCATATGGTATTA
GTAGCACCATATTGGAATGTAAATAACTGCACAGTATCACCGCTAGCTTTTAATTCTTTAGTA
TTAGTAGCACCATATTGGAATGTAAATTTTTGAAGTATATTATAAAGGCACAGTAACACGCCC
GTATTAGTAGCACCATATTGGAATGTAAATGCTTTAACTCTTAAAAAAAGATAAAGTTCTAAAT
TCGTATTAGTAGCACCATATTGGAATGTAAATAAGATGCAGCCAACGCACTTGGATATATGGC
TTTGGGTATTAGTAGCACCATATTGGAATGTAAATGGTAATGTAAGTAATTCTACAACCAATA
ATAGCAATGTATTAGTAGCACCATATTGGAATGTAAATATACAGAATACAAGATTATAGTTAGT
GGATATAGAAGTATTAGTAGCACCATATTGGAATGTAAATAATGAAGTCAAATAATAACATAC
CATTTTGTGCTCGTATTAGTAGCACCATATTGGAATGTAAATTTTAAATCTGGTTTATTTTTTA
CATTCTTCCAATCCGTATTAGTAGCACCATATTGGAATGTAAATAAGTGCGTTATTTACGCCTT
CTATATGTCCGAATACCGTATTAGTAGCACCATATTGGAATGTAAATTTTTCATTAGCAATTCC
TTTGTACTGCCATCTTCCGGTATTAGTAGCACCATATTGGAATGTAAATCCTAGTACGCCCAG
CATACCCAAAAAAGAACTACTTAAGTATTAGTAGCACCATATTGGAATGTAAATCTGGAAAG
AGGCAATAAAGCATTAGGAATAATAAAATGGTATTAGTAGCACCATATTGGAATGTAAATCTT
ACTAACACTTTCAGACCTAGTATTAAAATAATTTTGTATTAGTAGCACCATATTGGAATGTAA
ATGTGTTACATCTCCCAATTTCTCCTCATAATACTTTAAGTATTAGTAGCACCATATTGGAATG
TAAATGCTATAGCTAGTATAGTAGATACGTTGCGAGAATGGGTATTAGTAGCACCATATTGGA
ATGTAAATAATTAATATTGGCAGTATATGCTATACCATCTATAGCGTATTAGTAGCACCATATTG
GAATGTAAATTATCTAACTCAATATTTTCTTCTTTTACATCCTGTTTAGTATTAGTAGCACCATA
TTGGAATGTAAATAAATTAATAAAGATAGTAGGTTAAAGGGTATATTAGGTATTAGTAGCACC
ATATTGGAATGTAAATTTATAAAATTCTATTTCTAGTTCTTCTTGAGTATATGTATTAGTAGCAC
CATATTGGAATGTAAATGTCGACCCATTGGAGTTAGACAGATGGGATTTTCAGTATTAGTAGC
ACCATATTGGAATGTAAATTTAGCATCTATAATATTTACTTCTTTAATAGTTCTTGTATTAGTAG
CACCATATTGGAATGTAAATATATGTGATGATGAATTAGAGAAAGTGCTTGAAAGGTATTAGT
AGCACCATATTGGAATATAAATTGACTGTAACTATAGTATAAATAGGTAAATAGGGACGGTTA
ATGAATTTTTCAAAAGTATGATATAATAAAAACATAATTTACTTGATGAGGTATGATATGGATT
TTAATGAAACAATATATAATAGAATATTAAAATTTGTTAAGGAGAATCCCGATAGTGTTTATTT
GCCACAGGATTTTGATGAAGCGGGAAAGAAAGATGCATATTTTATTTTTAATGCGAAATGCG
GAACGGAAAAATTTGAAATTGAAAATAGCAATAATTTAATTAAGCTTATATCAAATTATTTAA
ATGATGAAGCTCAATATAATGATTTAATAGAGTATATTCATGAGTTTCCAATAATAGTATATTAT
TTTGAGTTTTGTAGAATACTTGAAATGCAA

So, I was able to find a few CRISPR regions. The next step is to find similar matches between the variable spacer sequences (unhighlighted regions) and potential phage genomes. Exact matches may occur, but may also be unlikely due to rapid mutation of phages.

I implemented the following for Nitrosomonas europaea, but repeated the process for all of the variable spacer sequences :




I did not take the entire sequence, but rather the middle 20-or-so nucleotides. I did this because nucleotides outside the repeated regions may be associated with the CRISPR structure instead of collected foreign DNA[6].
However, this did not yield any matches with Nitrosomonas europaea.
Next, I did the same for C. tetani. A potential match was found.
Finding this sequence in a genome the size of the phage's is very unlikely.

AATGAAGTCAAATAATAACATACCATTTTGTGCTC

AATAATAACATACCATTTT

Chance of occurring randomly in genome the size of organism's: $9.18 \times 10^{-7}$

Finding this sequence randomly in a genome the size of the phage's is very unlikely. It should be noted that other matches were found, but many were based on weaker matches that could have easilly happened by chance.

However, according to biobike this is a phage that infects Prochlorococcus, which is a cyanobacterium that lives in the ocean. As C. tetani is terrestrial, its unlikely the two came in contact with echother. The sceific sequence matches only this phage, and is located towads the end of a Phytanoyl-CoA dioxygenase gene. So, the reeason for this match isn't very clear, and could actually be due to coincidence (1 in 90 million can still occurr).

Next I looked back at Mycobacterium-tuberculosis-H37Ra. I noticed that this bacterium actually had associeted phages that had available gnomes in Biobike. I decided I should focus on this organism.

Using the method described above, I got the following four matches (Green are matching nucleotides)

Phage - ATCGACGCGAACCTGTCTGACGCG

Bacterium-TTCGAGCCGAACCCCGGTGACGCG

Chance of occurring randomly: $4 \times 10^{-6}$

Phage: Mycobacterium-phage-KBG

This is promising, as the phage is known to specifically infect mycobacteria.

Phage- GGAAGAGATCACGAATCCGGC

Phage- GGAACAGATCAGCCAACGGAC

Bacterium- GGACGTGATCAACGATCCGGC

Chance of occurring randomly: 2.5 x $10^{-4}$

These are Mycobacterium-phage-Kamiyu and Mycobacterium-phage-Tweety. This is good, as these both supposedly infect mycobacteria as well. Additionally, the differences seen are whole-codon changes, which is also promising (as frame-shifts are damaging but codon switches can have much smaller effects).

I also found;

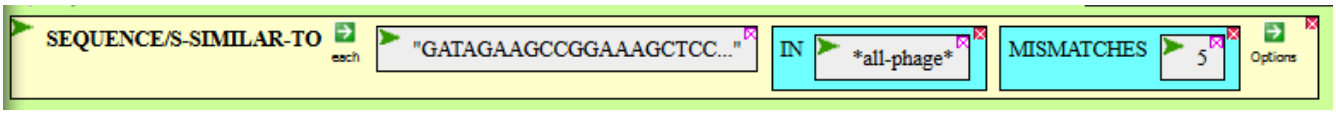
Bacterium- ACGAGGGCGCGGGGCACCCG

Phage- AAAAGGGCGCGGGGCAACCG
(Mycobacterium-phage-kikipoo)
Probability of occurring randomly in phage genome: 1 x 10-6

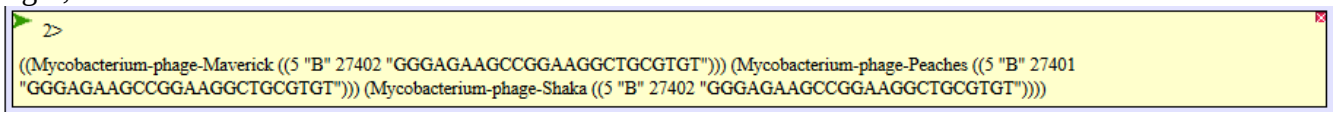
This was a very simila sequence, again occurring in a bacteriophage that attacks the genus of the bacteria whos CRISPR is being analyzed.

I decided to search some more.
Taking the 24 bp sequence
GATAGAAGCCGGAAAGCTCCGTGC

And searched while allowing for 5 mismatches.




I got;




Mycobacterium-phage-Maverick
Mycobacterium-phage-Peaches
Mycobacterium-phage-Shaka
Which all have the same matching string

Bacterium- GATAGAAGCCGGAAAGCTCCGTGC
Phage -GGGAGAAGCCGGAAGGCTGCGTGT

```
27701 GACGAGTTCGGGCTACCGGATACAACGGAGCCCCTGCTGGCCGTCCTGGTTGCGGGACGT
27641 GTCCACGTCGAACGAGTAACGCCAGACCTGGCGTCGATAGTTACAGCTCGGGAGGGACAC
27581 GTTGTGTCGCAGTACATGCAGAGGCTAACGCGCTTCTCTACTGCGATAGAGAGGACCTTC
27521 GCGGAGCTGCTCTCTACATCACCAGAGCCCCCTGCGGAGACTGCTCGAAGCTGATCGACG
27461 CCGCCGGCATCGAGCGAGTGGTGTACCCGTTCGAGTGGGAGAAGCCGGAAGGCTGCGTGT  conserved hypothetical protein
27401 GTGCTGGGCCCACCGACCCGTACCACGGGTGGTGCGATCCGTGCAGCAAGCTCCCGAAGA  Maverick.Mave-0036 (27376 -> 27900)
27341 TCGTGGTGCAAGAGCAGGACCGCTGGGAGCGGAACGAGTACACCGTCCGGATGGAACCGG  Maverick.Mave-0035 (27211 <- 27405)
```

These are all located inside a hypothetical protein.

I did this again with:
GTTACCCACGCCGATTTACTGGCC

```
8>
((Mycobacterium-phage-Vix ((5 "F" 26138 "GTTGCCGACGACGATCTTCTGGCC"))) (Mycobacterium-phage-JHC117 ((5 "F" 26054
"GTTGCCGACGACGATCTTCTGGCC"))) (Mycobacterium-phage-Pukovnik ((5 "F" 45895 "GCTACCCACGCGCATATACTCGCC"))) (Mycobacterium-phage-Rockstar ((5
"F" 25324 "GTTGCCGACGACGATCTTCTGGCC"))) (Mycobacterium-phage-Microwolf ((5 "F" 26057 "GTTGCCGACGACGATCTTCTGGCC"))))
```

Again, all mycobacteriophages.
Mycobacterium-phage-Vix
Mycobacterium-phage-JHC117
Mycobacterium-phage-Pukovnik (with a slightly differrent matching sequence)
Mycobacterium-phage-Rockstar
Mycobacterium-phage-Microwolf

Phages- GTTGCCGACGACGATCTTCTGGCC
Phage Pukovnik- GCTACCCACGCGCATATACTCGCC
Bacterium- GTTACCCACGCCGATTTACTGGCC

For the phage vix, the match is seen here-

```
Select org/contig  Start  Prev  Next  End
26078 AACCATCGCCGCGACGTGTGGCGGCACGGTCACCGGACGCTTCGAGCGGACGGTCTTGGT  Integrase
26138 GTTGCCGACGACGATCTTCTGGCCGACGCGGGCCGCGCCCCGGCGCACCCGGAACTTCAT
26198 CGTCACGCCGTCGTCCATGATGTCCTTCCGGCGAATCTCGATCAGCTCCCCGAACCGCAG
26258 GCTCGTCCACGCGAGGAGGTAGACGGCCACCCGGTAGTGCTCATGGACCTCGGCTGCGAC
26318 CGTCTCCAATTCCTGGGGCGTGAGAGCCTCCACGTCGCGCTCGTTGGGGGCCTTCTGCTC
26378 GATCCGGCACGGGTTCTCCGAGAGCAGCTTGTCCTCGACGGCGGTGTTCATGACGGCCCG
26438 GAGGATGTTGTAGGCGTGTCGACGCGCTGTCGGGTGCTGATGCCCCATGCCGGCCCACCA
26498 CGCCCGGACGAGGGCTGGCGTCATCTCGGCGACCGCCGTGTCGCCCAGCACCGGGTAGAT
26558 CCGCTTCCTGGCGTGGGTCTTGTACAGCTCCCGCGTCCCCTCCGCGAGGTCGCGCTCAGT
```

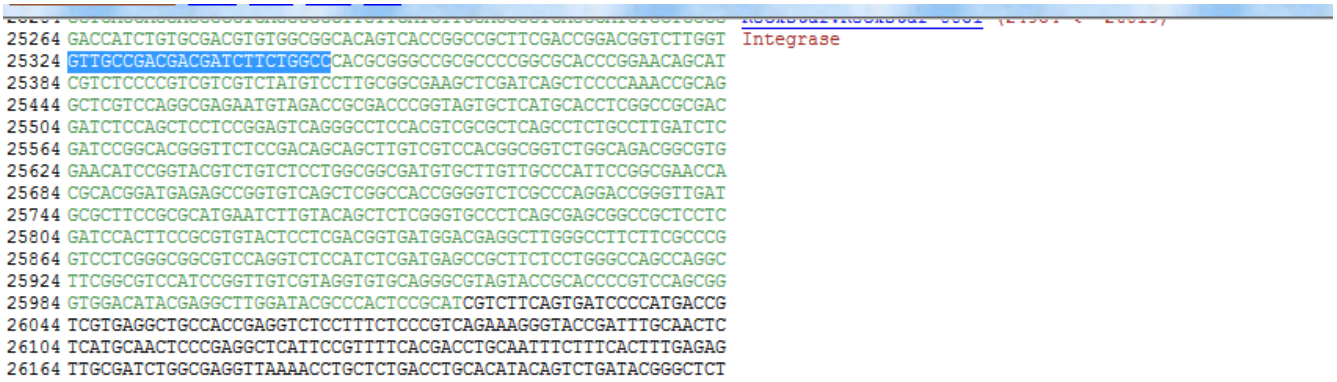
Located at an integrase.

For JHC117,

```
25994 GACCATCGCCGCGACGTGTGGCGGCACGGTCACCGGACGCTTCGAGCGGACGGTCTTGGT  Integrase
26054 GTTGCCGACGACGATCTTCTGGCCGACGCGGGCCGCGCCCCGGCGCACCCGGAACTTCAT
26114 CGTCACGCCGTCGTCCATGATGTCCTTCCGGCGAATCTCGATCAGCTCCCCGAACCGCAG
26174 GCTCGTCCACGCGAGGAGGTAGACGGCCACCCGGTAGTGCTCATGGACCTCGGCTGCGAC
26234 CGTCTCCAATTCCTGGGGCGTGAGAGCCTCCACGTCGCGCTCGTTGGGGGCCTTCTGCTC
26294 GATCCGGCACGGGTTCTCCGAGAGCAGCTTGTCCTCGACGGCGGTGTTCATGACGGCCCG
26354 GAGGATGTTGTAGGCGTGTCGACGCGCTGTCGGGTGCTGATGCCCCATGCCGGCCCACCA
26414 CGCCCGGACGAGGGCTGGCGTCATCTCGGCGACCGCCGTGTCGCCCAGCACCGGGTAGAT
26474 CCGCTTCCTGGCGTGGGTCTTGTACAGCTCCCGCGTCCCCTCCGCGAGGTCGCGCTCAGT
```

Also at the integrase protein.

Rockstar:

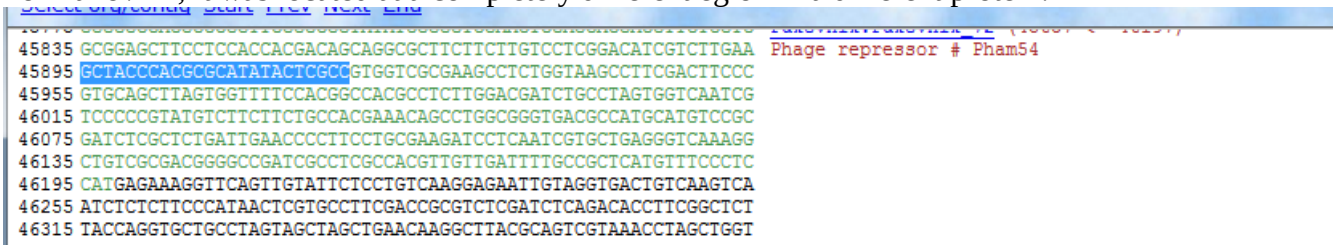



Also at itegrase.

Microwolf:


Also at integrase

For Pukovnik, it was located at a completely different egion in a different protein.




A phage-repressor gene.

It is interesting that both viruses contain the matching sequence yet do not appear to be related. Still, it is reasonable to assume that
Vix, JHC117, Rockstar and Microwolf are all related to each other closely, as they show highly homologous DNA sequences.

I again searched with:
TGCGCGCACAACGCATCCGCCATCCA

One match:

17> ((Mycobacterium-phage-Wildcat ((5 "B" 36894 "TCCGCGCGCAACACATCAGCCAGCCA"))))

Again, a mycobacterium phage.
Bacterium- TGCGCGCACAACGCATCCGCCATCCA
Phage- TCCGCGCGCAACACATCAGCCAGCCA

```
37373 AATAGCGCCAGGGTCCCACTTGCCCTGCGCGCGCTCGCCCCACTCACGGTGCGAGATACA
37313 CCTATCGGCACGCAACCCCAGCTTGCGGAGAATAGCCGCATGCACCTTCACAGTTGCGTC
37253 GTACTGCACATCTGGCCAGCCCTCACGGTGCGGAGCATCCTTGCGAGGCAGGATCGCCAC
37193 CTCGACGCCGATCATCACCGGGTTCGCGTTGTTTGTAGGCAGGCCAGGATAATCGCCACG
37133 ACCAGCGTGGTTCGCCTTACCGATACCACACACCCACACGTCGCCGTTCGGCTTGATCAG
37073 AATGTGGGCAGCCAAACCCAAAGTCGGGTGGAAGGCAATGCCCTCCGGAGTCTCGTTCGG
37013 AGAACCAGTGTGGTGGAACATGACCCCCCAGAAAGTTCCCTGGTCGCCCTGGCCACGGTT
36953 CTGCCAACCATCCACCTCGAACACGCGAAGACCCTCCGCGCGCAACACATCAGCCAGCCA
36893 CCACGGGTCACCAGTGAACCCAGGCTCAGGAGTAGTGGCAGGAGTGTAGGACCCATCAAA Phage endolysin
36833 CGGGATGGCCAGCACGCGATTCCATCGAGCCTGACGATCCGTCCAGCCATTCGGAAGCTG Wildcat.Wildcatp49 (36288 -> 37796)
```

This is a page endolyain gene.
More searching with:
GGGGGACTGTGGACGAGTTCGCG

19>
((Mycobacterium-phage-DS6A ((5 "F" 35433 "GGGTGACGGTGGAGGAGGTTGCG"))) (Mycobacterium-phage-Pari ((5 "B" 37214 "GGGGGACCAAGGGCGAGTTCGAG"))) (Mycobacterium-phage-Backyardigan ((5 "F" 31589 "GGGAGACGGTGGATGGGTTCACG"))))

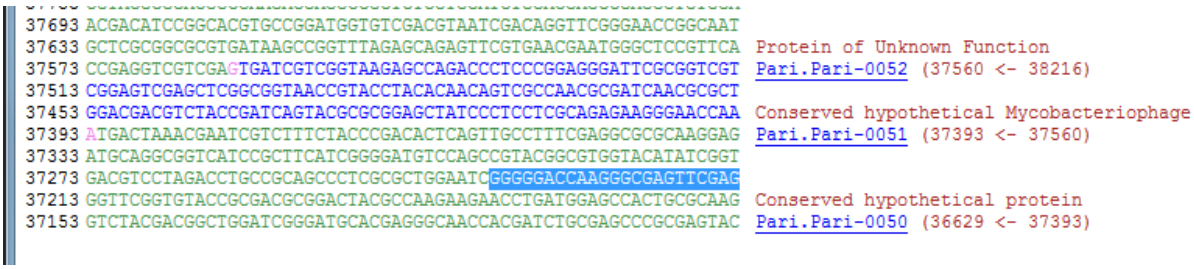
Three matches. All mycobacteria. DS6A, Pari, Backyardigan. All slightly different.

Bacterium: GGGGGACTGTGGACGAGTTCGCG
DS6A: GGGTGACGGTGGAGGAGGTTGCG
Pari: GGGGGACCAAGGGCGAGTTCGAG
Backyardigan: GGGAGACGGTGGATGGGTTCACG

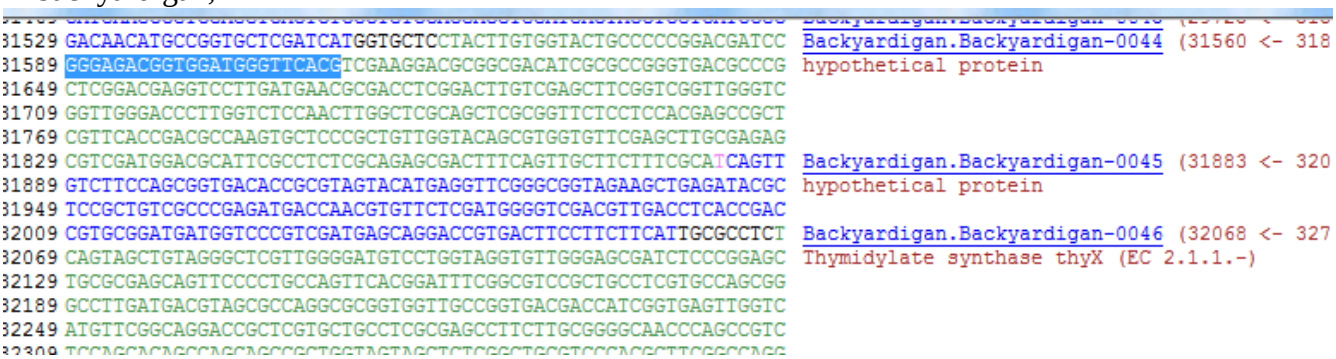
```
35373 TGCACGTGAAGCTGACGGCGGCGACGCTGACCGAGTACCACCGGGTTGCTGCGGCGGCGG hypothetical protein
35433 GGGTGACGGTGGAGGAGGTTGCGCGGGAGGCGATTGCGGAGCGGTTCGCGGTGATGGTGA
35493 AGGCGCAGGCGGGCGGCCGGCGTCGGGGCGGTAAAGCGGTGGCATGATGGCCGCCACATT DS6A.DS6A-0051 (35536 -> 36210)
35553 GGACCGCCGGCAGTGCCCGGGGTGTGGCGAGGGCCTGCCGGCCGACGCCCACCCGCTGCG hypothetical protein
35613 GAAGTGGTGCAGTGAGCGGTGCCATTCTCGGGCACGAGACGCCCGGCCTGATCGGCGAGC
35673 GAAACGTAAGGCTCGGAAGCGGGAACGGCACCGGGAGCGCATGGCGACGGATCAGGCATA
35733 CAAGGATCGACACCGGGCCAAGAGGCGGGCGCAGAAGCAGCGGCAGCGGGCGGCGAAAGT
35793 GGCCGCCATGCCGCTGCGGCAGTGCGCATGGTGCGGCGAGTCGTTGCGGGCCGGCGCCCC
```

In DS6A this occurrs in a hypothetical protein

pari, this also occurrs in a hypothetical protein.



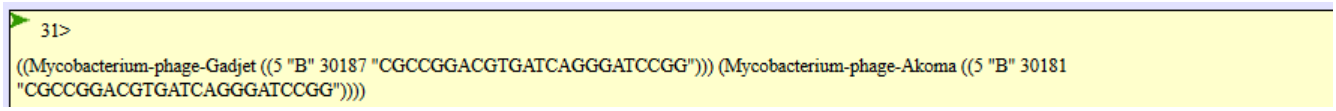

in backyardigan,




Another hypothetical protein.

Another search:

CGCGGGAAGAGATCACGAATCCGG



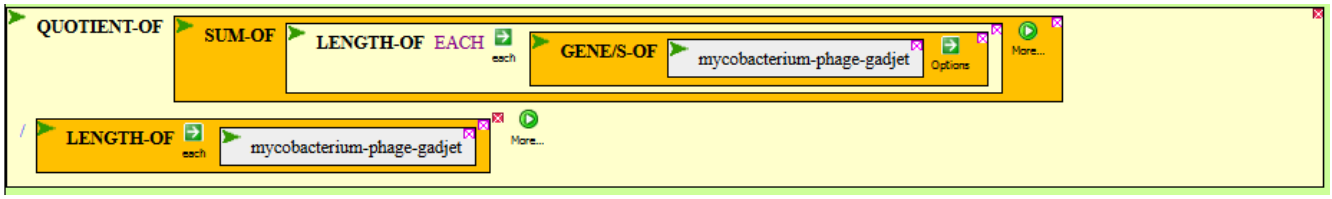

Two more mycobacteriophages.

Bacterium: CGCGGGAAGAGATCACGAATCCGG
Gadjet and Akoma: CGCCGGACGTGATCAGGGATCCGG


Located on a phage protein.

Out of the 8 phages looked at so far, allmatching items have been located on a gene.

To see if this was odd or not, I decided to look at what roportion of the paghes' genomes were coding regions.

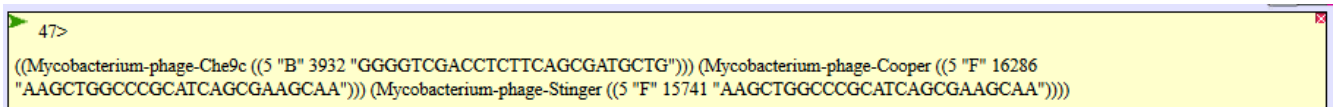


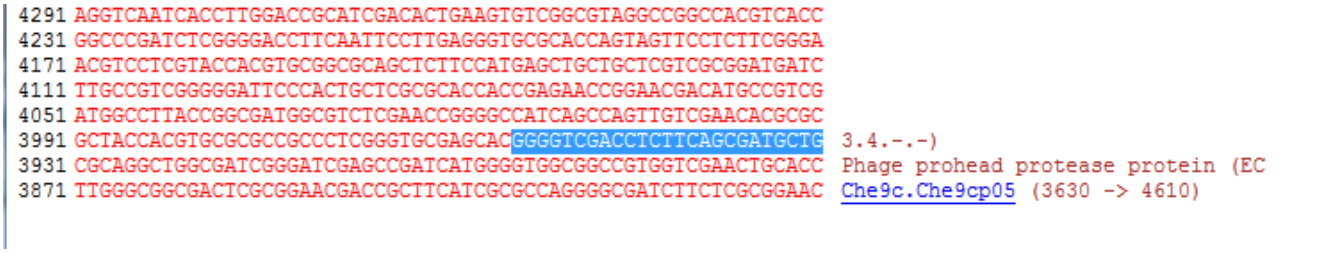

Gadjet – 95%
DS6A - 94%
Pari- 90%
Backyardigan – 91.8%
Wildcat- 92.2%
Vix – 89.9%
Pukovnik – 93.1%
Maverick – 91.9%

Average: 92.2% coverage.

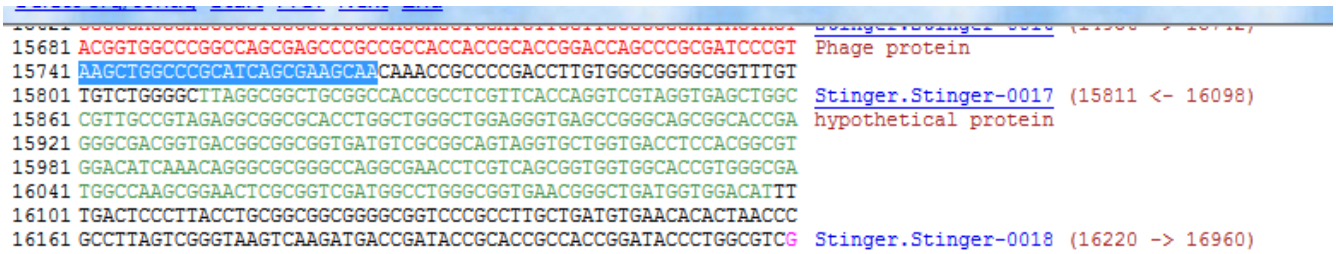
So, there's about a fifty percent chance ($92.2^{8}$) that I would find the sequences exclusively on genes. More searching.

GAGCTGGACCGCATCAGCGATGCTG




Prohead protein. Coding region: 94.7%




For the mycobacteria phage stringer, the match actually fell (mostly) between two declared genes. Coding region is 94.9%.

More searching...

Another match:

94.9%

So, there was one instance of a non-gene match and 10 instances of gene matches. This is 91% actual instances of the match landing on a gene. The avergage gene coverage for the viruses so far is 92.9%. This is reasonably close to what would be expected.

Clearly more data could be obtained and would likely be beneficial in elucidating a further pattern. However, based on the obtained reuslts, it seems reasonable to conclude that there is a fair chance there is no correlation between coding/noncoding viral DNA and the variable spacer regions in CRISPRs. Still, it was interesting to find a number of matches to known mycobacterial phages, as this likely indicates previous contact in the organisms evolutionary history.

I think I was able to answer my question, at least in a preliminary sense in that a lot more research would be required to really answer the question but this was a decent start. I found very little relationship between coding vs non coding regions in phages and CRISPR regions in bacteria. I was also able to identify quite a few phages that likely infect one particular mycobacteria. Some of the phages seemed related, others didn't, and some even contained highly similar sequences even though they otherwise seemed very much unrelated (what was very surprising – perhaps they picked up similar DNA during a lysogenic phase, or some other mechanism for horizonal gene transfer.

Future resarch into this would benefit from further automation of the process, although quite a few of the key programs I used took between 10-50 secods for biobike to execute. Further automation may take long periods of time for the computer to run. Still, more data could construct a better picture and determine if there is infact a relationship between selection of variable sequences and the region of the phage genome it comes from. This could actually be useful, as CRISPRs are rapidly being intergrated

into biotechnology purposes, and further understanding could benefit the field.

References

[1] Hans-W Ackermann. 2011. Bacteriophage taxonomy. Under the microscope. 90-94 http://journals.cambridgemedia.com.au/UserDir/CambridgeJournal/Articles/11%20ackermann244.pdf
[2]Wommack, K. E.; Colwell, R. R. 2000. Virioplankton: Viruses in Aquatic Ecosystems. Microbiology and Molecular Biology Reviews 64(1): 69–114
[3]Horvath, P., Barrangou, R. January 2010. CRISPR/Cas, the Immune System of Bacteria and Archaea. Science. 327(5962): 167-170
[4]Haft, D. H.; Selengut, J.; Mongodin, E. F.; Nelson, K. E. 2005. A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. PLoS Computational Biology 1(6): e60
[5]Barrangou, R.; Fremaux, C.; Deveau, H.; Richards, M.; Boyaval, P.; Moineau, S.; Romero, D. A.; Horvath, P. (2007). CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science 315(5819): 1709–1712.
[6]Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, et al. 2008. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. Journal of Bacteriol 190: 1401–1412