Donald Ta
BNFO301

Journey into Clustered Repeats of Bacterial Genomes of the Pseudomonas Order

Introduction:

In Biology we are usually just thrown information about how processes work and how certain actions of an organism are regulated. Very rarely do we actually get a chance to explore deeper into the structures that cause these points of regulation. Palindromic sequences are very abundant in genomes of bacteria. These palindromic sequences can range from any size and are often important in specific steps of gene transcription, and protein binding sites. I wanted to get a bit deeper than that, so I looked at clustered repeats.

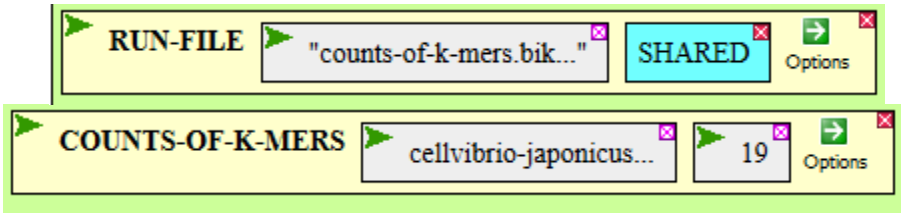
Clustered repeats is a very broad and general term for sequences of nucleotides that are often repeated throughout the genome of a bacterium. This whole group is composed of sequences identified such as repetitive extragenic palindromic sequences (REPs), clustered regularly interspaced short palindromic repeats (CRISPRs), bacterial interspersed mosaic elements (BIME), and enterobacterial repetitive intergenic consensus sequences (ERICs) just to name a few. The history of how all of these clustered repeats came about in our modern day bacteria in such large numbers is still widely unknown. Some of these repeats have been more researched such as CRISPRs, while some still have lots of mystery surrounding their purpose and existence such as REPs.

CRISPRs in particular tend to diverge very quickly among species. [3]. REPs are usually very species specific but can share similar repeated elements. [1]. Both of these sequences are capable of acting as regulatory sites for a variety of functions such as transcriptional terminators, specific protein binding sites, and structure stabilizers. This is because many of these types of clustered repeats have a very unique trait in that they are palindromes and are capable of forming into stem-loop-structures or hair-pin structures by pairing up the nucleotides to their compliments on the same strand. [2].

These clustered repeats have been found extensively in Enterobacteriaecae which encompasses organisms such as E. *coli* and its many strains. I first started off with a known case of clustered repeated called REPs. It was the first time these REPs were found in bacteria outside of Enterobacteriaecae.[1]. Within the Pseudomondas order I was able to find the clustered repeat that they found in P. *putida* KT2440. However, as my journey into finding these clustered repeats progressed, my findings were quite different than I expected. My ultimate question was whether these repeats shared any conserved regions among the rest of the species in the same order?

Methods and Results :

In PhAnToMe/BIOBike, I first loaded the COUNTS-OF-K-MERS function into BioBike. I first used the function to find the clustered repeat found in P. *putida* KT2440 by looking for the most repeated sequence and then using MATCHES-OF-PATTERN to find the sequence itself in the genome. I was able to find the sequence and note that it was outside of genes and that it repeated 810 times throughout the genome. So after figuring out that my method worked for finding a clustered sequence, I used the COUNTS-OF-K-MERS function again to find the most repeated sequences throughout the genome of C. *japonicus*. I used this organism because it is in the same order as the P. *putida* and was readily available in BioBike. I wanted to determine if the conserved regions in P. *putida* would be found in any repeats found in C. *japonicus*. I chose a window size of 19 because these clustered repeats are generally from 20 to 40 nucleotide repeats so I felt that 19 would be a good number to encompass a broad range of possibilities of the repeats and also because the program wouldn't work for any number past 20. I looked at the top results and found them to repeat 72 times in the genome of C. *japonicus* which can be seen in Figure 1.



COUNT	WORD
73	ACGCGCTCGCGAAGAGCGC
73	CACGCGCTCGCGAAGAGCG
73	CGCGCTCGCGAAGAGCGCG
73	CGCTCGCGAAGAGCGCGAC
73	GCGCTCGCGAAGAGCGCGA
72	AATCCACGCGCTCGCGAAG
72	ATCCACGCGCTCGCGAAGA
72	CAATCCACGCGCTCGCGAA
72	CCACGCGCTCGCGAAGAGC
72	GTTTCAATCCACGCGCTCG
72	TCAATCCACGCGCTCGCGA
72	TCCACGCGCTCGCGAAGAG
72	TTCAATCCACGCGCTCGCG
72	TTTCAATCCACGCGCTCGC

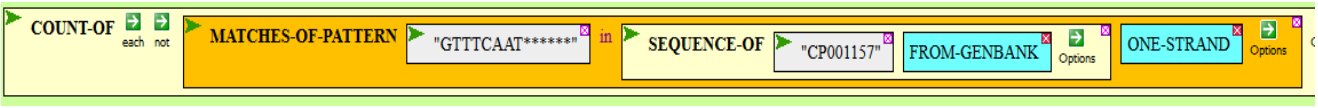
Figure 1: The COUNT-OF-K-MERS output for C. *japonicus* showing the most highly repeated sequences in the genome. The number on the left under count indicates how many times the sequence under word came up within the genome.

I ran a MATCHES-OF-PATTERN of the sequence circled in red in Figure 1 and found there to be the same 72 repeated sequences in the genome of C. *japonicus*. The chances of a specific 19 nucleotide sequence occurring the genome size of C. *japonicus* is about 3.63×10^{-12} in this genome size because the ratios of A, G, C, and T are all about ¼. So it was highly unlikely that this sequence came about by chance. The results of the first search can be seen in the Figure 2 Appendix.

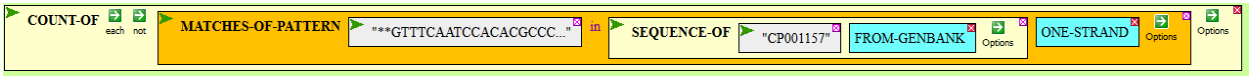
Something that was interesting about this sequence was that they were palindromic sequences with a 7-nucleotide linker. It actually looks like it could form a stem-loop-structure or even a hairpin structure because of this. I didn't realize until I looked more closely that the sequence being repeated actually extended 11-nucleotides downstream(to the left) and 2 nucleotides upstream (to the left) with those sequence parts being conserved 100% in the other 71 repeats. At first I also thought that this was an REP because the sequence had the general structure of an REP, two tandem inverted repeats separated by a linker with most of these repeats

outside of genes (extragenic). However, what caught my eye was that every single one of these repeats was separated by about 34 nucleotides and were only located in this one specific portion of the genome. The full repeated sequence can be seen in Appendix Figure 3.

I did a BLASTN search of this repeated sequence (found in Appendix Figure 3) with all of the organisms in the Pseudomonas order in order to determine if this sequence was conserved in any other species. I picked to run with the somewhat similar matches just because I know that many of these clustered sequences are species-specific. One of the organisms that I was interested in was A. *vinelandii*, however the organism's genome is not complete on BioBike so I looked up the specific strain that I thought had a lot of matches to the sequence I was looking for. Then I performed a MATCHES-OF-PATTERN to see if I could get the same matches that I did on the blast search.

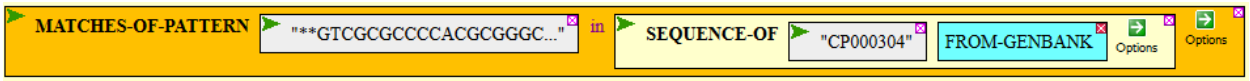


It was interesting to see that there were actually many matches (47 to be exact) to this sequence in this organism each with a conserved 8 nucleotide sequence similar to the sequence found in C. *japonicus*. All of the repeats also ranged between 50-60 nucleotides in distance from each other. However, there doesn't seem to be a palindromic sequence within this repeat. Since all of these matches seemed to be perfect matches, I wanted to expand it and see how long this repeat actually was. Through an unpractical, but working way, I just placed the rest of the repeat in and extended the "*" until it stopped repeating perfectly. The "*" denotes an unspecified nucleotide. When the function iterates through the genome, only the matches containing the specific nucleotides written will be found and all of the "*" can be any other nucleotide.



The perfect repeat was 32 nucleotides long with a gap between each repeat ranging from 30 to 35 nucleotides on average. Again, it still contained the conserved sequence "GTTTCAATCCAC" found at the beginning of the sequence in C. *japonicus* as well. The results of this search can be found in Appendix Figure 5.

Most of the Pseudmonas did not seem to have good matches except for P. *stutzeri* according to the blast output. However, when I tried to search for it at first there didn't seem to be any matches. I relooked at it again and I realized that it was actually located on the opposite strand.



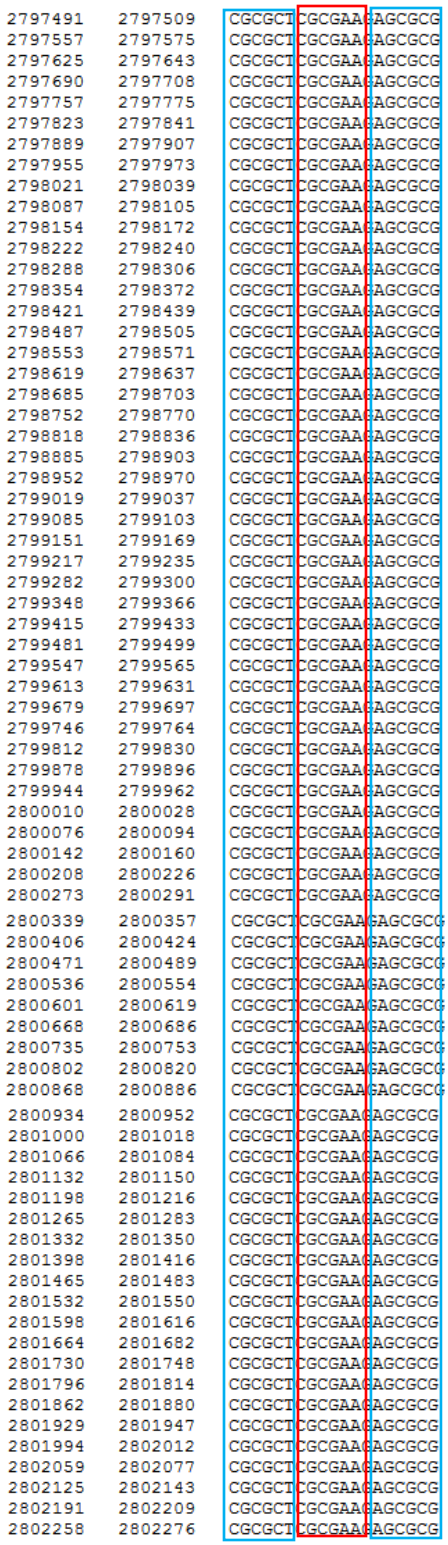
So after taking the complement of the sequence I wanted to try to see if I could locate these repeats. What surprised me was that the repeats were exactly the same as the ones found in C. *japonicus* except just on the opposite strand.

I wanted to see if this sequence occurred outside of the Pseudomonas so I went up one level to see if it occurred in any organisms in the Moraxellaceae. I did the MATCHES-OF-PATTERN on Acinetobacter-sp-ADP1 as well as psychrobacter-sp-prwf-1 and could not find the sequence or its compliment. So then I wanted to determine if they had these repeats of their own that I could compare. So again I did COUNTS-OF-K-MERS on both and looked at the most repeated sequences. What was interesting about the Acinetobacter-sp-ADP1 was that it went through many different genes with unknown functions. This repeated sequence can be found in Appendix Figure 8. Acientobacter-sp-ADP1 had about 32 nucleotides separating between each of the repeats and occurred 90 times, which was a lot more frequent than the previous organisms that I looked at. None of the repeats seemed to be significantly similar to anything found in psychobacter-sp-prwf-1.

Conclusion and Future Plans:

Based on the data that was collected, I realized that the original clustered repeat that I was looking for did not share any conserved regions in the clustered repeats found in C. *japonicus,* A. *vinelandii*, or P. *stutzeri*. However, I did find conserved regions present in the Pseudomondas order. I also believe that the repeated sequence found in C. *japonicus* that I had found first is actually a CRISPRs and not an REP. I believe this is because of the way it is oriented one after another and also because the sequence happens to lie next to a Cas(CRISPR associated sequence). [3]. The same thing could be said about the repeated sequence in A. *vinelandii*, however I did not see a Cas near the repeated sequence in P. *stutzeri*. Another interesting thing that I observed was that all of the repeats that I found were also exactly 32 nucleotides in length. The lengths of the gaps between each repeat were different, but I found it quite interesting that the size of the repeat itself was conserved. This could mean they share a similar origin and could definitely point to how the clustered repeat came about in the genomes of their respective bacterial species. I hypothesized that there would be conserved regions of clustered repeats within the Pseudomonas order and I was able to observe some of those conserved regions. However, I also wanted to know whether or not these regions would be conserved specifically in this group of organisms or could it be found elsewhere. So when I looked at the organisms in Moraxellaceae, I did not find a repeat that was conserved between the two organisms nor did I find a similarity with those found in Pseudomondas. It has been said that most of these repeated sequences don't share similarities to others and are species specific. However, I was able to find at least some conserved regions of repeated sequences within the Pseudomondas order.

In the near future, I would hope to determine where these sequences come from. It has been stated in past studies that these could have come from transposases or even from phages. If I were to continue this research, I would look into what phages, if any, would have similar sequences to the one found in C. *japonicus*. I feel like this would be able to help the scientific community be one step closer to locating another origin of these repeated sequences in bacterial genomes.



2797491	2797509	CGCGCTCGCGAAGAGCGCG
2797557	2797575	CGCGCTCGCGAAGAGCGCG
2797625	2797643	CGCGCTCGCGAAGAGCGCG
2797690	2797708	CGCGCTCGCGAAGAGCGCG
2797757	2797775	CGCGCTCGCGAAGAGCGCG
2797823	2797841	CGCGCTCGCGAAGAGCGCG
2797889	2797907	CGCGCTCGCGAAGAGCGCG
2797955	2797973	CGCGCTCGCGAAGAGCGCG
2798021	2798039	CGCGCTCGCGAAGAGCGCG
2798087	2798105	CGCGCTCGCGAAGAGCGCG
2798154	2798172	CGCGCTCGCGAAGAGCGCG
2798222	2798240	CGCGCTCGCGAAGAGCGCG
2798288	2798306	CGCGCTCGCGAAGAGCGCG
2798354	2798372	CGCGCTCGCGAAGAGCGCG
2798421	2798439	CGCGCTCGCGAAGAGCGCG
2798487	2798505	CGCGCTCGCGAAGAGCGCG
2798553	2798571	CGCGCTCGCGAAGAGCGCG
2798619	2798637	CGCGCTCGCGAAGAGCGCG
2798685	2798703	CGCGCTCGCGAAGAGCGCG
2798752	2798770	CGCGCTCGCGAAGAGCGCG
2798818	2798836	CGCGCTCGCGAAGAGCGCG
2798885	2798903	CGCGCTCGCGAAGAGCGCG
2798952	2798970	CGCGCTCGCGAAGAGCGCG
2799019	2799037	CGCGCTCGCGAAGAGCGCG
2799085	2799103	CGCGCTCGCGAAGAGCGCG
2799151	2799169	CGCGCTCGCGAAGAGCGCG
2799217	2799235	CGCGCTCGCGAAGAGCGCG
2799282	2799300	CGCGCTCGCGAAGAGCGCG
2799348	2799366	CGCGCTCGCGAAGAGCGCG
2799415	2799433	CGCGCTCGCGAAGAGCGCG
2799481	2799499	CGCGCTCGCGAAGAGCGCG
2799547	2799565	CGCGCTCGCGAAGAGCGCG
2799613	2799631	CGCGCTCGCGAAGAGCGCG
2799679	2799697	CGCGCTCGCGAAGAGCGCG
2799746	2799764	CGCGCTCGCGAAGAGCGCG
2799812	2799830	CGCGCTCGCGAAGAGCGCG
2799878	2799896	CGCGCTCGCGAAGAGCGCG
2799944	2799962	CGCGCTCGCGAAGAGCGCG
2800010	2800028	CGCGCTCGCGAAGAGCGCG
2800076	2800094	CGCGCTCGCGAAGAGCGCG
2800142	2800160	CGCGCTCGCGAAGAGCGCG
2800208	2800226	CGCGCTCGCGAAGAGCGCG
2800273	2800291	CGCGCTCGCGAAGAGCGCG
2800339	2800357	CGCGCTCGCGAAGAGCGCG
2800406	2800424	CGCGCTCGCGAAGAGCGCG
2800471	2800489	CGCGCTCGCGAAGAGCGCG
2800536	2800554	CGCGCTCGCGAAGAGCGCG
2800601	2800619	CGCGCTCGCGAAGAGCGCG
2800668	2800686	CGCGCTCGCGAAGAGCGCG
2800735	2800753	CGCGCTCGCGAAGAGCGCG
2800802	2800820	CGCGCTCGCGAAGAGCGCG
2800868	2800886	CGCGCTCGCGAAGAGCGCG
2800934	2800952	CGCGCTCGCGAAGAGCGCG
2801000	2801018	CGCGCTCGCGAAGAGCGCG
2801066	2801084	CGCGCTCGCGAAGAGCGCG
2801132	2801150	CGCGCTCGCGAAGAGCGCG
2801198	2801216	CGCGCTCGCGAAGAGCGCG
2801265	2801283	CGCGCTCGCGAAGAGCGCG
2801332	2801350	CGCGCTCGCGAAGAGCGCG
2801398	2801416	CGCGCTCGCGAAGAGCGCG
2801465	2801483	CGCGCTCGCGAAGAGCGCG
2801532	2801550	CGCGCTCGCGAAGAGCGCG
2801598	2801616	CGCGCTCGCGAAGAGCGCG
2801664	2801682	CGCGCTCGCGAAGAGCGCG
2801730	2801748	CGCGCTCGCGAAGAGCGCG
2801796	2801814	CGCGCTCGCGAAGAGCGCG
2801862	2801880	CGCGCTCGCGAAGAGCGCG
2801929	2801947	CGCGCTCGCGAAGAGCGCG
2801994	2802012	CGCGCTCGCGAAGAGCGCG
2802059	2802077	CGCGCTCGCGAAGAGCGCG
2802125	2802143	CGCGCTCGCGAAGAGCGCG
2802191	2802209	CGCGCTCGCGAAGAGCGCG
2802258	2802276	CGCGCTCGCGAAGAGCGCG

This was just the original MATCHES-OF-PATTERN search that I used to figure out where in the genome the repeated sequence was located. The reason why I searched for this specific sequence at first was because I was originally looking for two inverted repeats with a variable-sized linker in between.

As you can see in the figure to the left, the palindromic parts are circled in blue while the linker is circled in red.

APPENDIX: Figure 2: C. *japonicus* MATCHES-OF-PATTERN output

Figure 3: C. *japonicus* Sequence

```
2797546 GGAGGAGTTTGTAAGAGAGCAGGAGAGGGATAGGGTAATAGCTGCCAATTAAATACCACT
2797606 ACTCACGCGCTCGCGAAGAGCGCGACACTGAAACTGGCTCGCACGGCTCGCGCAGTACAA
2797666 GTTTCAATCCACGCGCTCGCGAAGAGCGCGACTGTAATCAACGACCGGGCTTTCTGCTCG
2797726 ATACCCTCGTTTCAATCCACGCGCTCGCGAAGAGCGCGACGCAGCGGCAGATGGTTTCTT
2797786 CCTGGCCGGTTTTGTTTCAATCCACGCGCTCGCGAAGAGCGCGACCCGGGAGCGCCGTTG
2797846 GGCGCGGGACCCCGTGAGCGGTTTCAATCCACGCGCTCGCGAAGAGCGCGACTGGGTTTA
2797906 ACTGCTGGTGCAGATGCTCATTGAGTGTTTCAATCCACGCGCTCGCGAAGAGCGCGACCG
2797966 ATGCCGCGTAAGCGAACCTTTACCCTTGAGCAGTTTCAATCCACGCGCTCGCGAAGAGCG
2798026 CGACCGGCGCTGGTAGTCGTACAGGGGCAATGCTGGCTGTTTCAATCCACGCGCTCGCGA
2798086 AGAGCGCGACGCATAACCAGCTGCCAACCTTTGAGGTGTTATATGTTTCAATCCACGCGC
2798146 TCGCGAAGAGCGCGACGCACCCGCTCGGCATCGGTACCAACAATTGTTGAGTTTCAATCC
2798206 ACGCGCTCGCGAAGAGCGCGACCACCTGTGGTGGGCTAGATGCGACTGCGGGAACGAGTT
2798266 TCAATCCACGCGCTCGCGAAGAGCGCGACGTCCGCGAGGATCTGCGCTGTACGGGTGTAG
2798326 TTCTCGTTTCAATCCACGCGCTCGCGAAGAGCGCGACTTTTAGACTGCAACGCAAAACAA
2798386 ATTTGCGTGGAGTTTCAATCCACGCGCTCGCGAAGAGCGCGACCCGCCAAGCAATCCACG
2798446 TTTGTCACCGAATATCTGTTTCAATCCACGCGCTCGCGAAGAGCGCGACGGACAACATCC
2798506 GGTGCTTTCAGCTTTCATCCACAGGTTTCAATCCACGCGCTCGCGAAGAGCGCGACTTCT
2798566 TGATGCTTTACCAGGGCCACTTTGCTCCCTGTTTCAATCCACGCGCTCGCGAAGAGCGCG
2798626 ACCATTAAATTCAGGGCAGAGAGCAACACCACCTATGTTTCAATCCACGCGCTCGCGAAG
2798686 AGCGCGACTCGATCAATAGATAGGTTAACATCTCGCCGAGTCGTTTCAATCCACGCGCTC
2798746 GCGAAGAGCGCGACCGCTGCGCTTTTTTGCCCTCGGCAAAATTAGCCGGTTTCAATCCAC
2798806 GCGCTCGCGAAGAGCGCGACCCCTTGGCGCGCTTTATAGACTACTCGACCTGTAGGTTTC
2798866 AATCCACGCGCTCGCGAAGAGCGCGACAACCGTATCACTGTCGGGTAGTCGTTTTTATTA
2798926 CGTTTCAATCCACGCGCTCGCGAAGAGCGCGACCAATCATGGCGGCTGCCAATAATCAAT
2798986 TAACGCCTGTTTCAATCCACGCGCTCGCGAAGAGCGCGACGGTATCAAGCCGCTGGCAAA
2799046 TGCTGCAAAAGCTGTGTTTCAATCCACGCGCTCGCGAAGAGCGCGACTATCCAATTTCGA
2799106 TTTACTCCGACCCATACATGAAGTTTCAATCCACGCGCTCGCGAAGAGCGCGACTTCGGG
2799166 TTAATACCGCACCGACAGATATACCGGTGTTTCAATCCACGCGCTCGCGAAGAGCGCGAC
2799226 GCGCTTTAATGATTCTACGCGTGACTGATATGCAGTTTCAATCCACGCGCTCGCGAAGAG
2799286 CGCGACTGCGCTACCCAGGTGAGCCGTCTGCACTCACCTGGTTTCAATCCACGCGCTCGC
2799346 GAAGAGCGCGACTTATAATCCGCCGTGCCTGTCTCCAGGTAGACCGTTTCAATCCACGCG
2799406 CTCGCGAAGAGCGCGACACTTTAGCGATGGCAACCAGAACATCAACAACAAGTTTCAATC
2799466 CACGCGCTCGCGAAGAGCGCGACGGCAATTCCTCTACATGATCGATGACGTACCTCCTGT
2799526 TTCAATCCACGCGCTCGCGAAGAGCGCGACTTGTGCGCGTTTACGCTCAGCGGCAGACAT
2799586 GGCAGTTTCAATCCACGCGCTCGCGAAGAGCGCGACCAAAAGTGTTTAAGGGGTAGGGTA
2799646 TGTTTATTCTGTTTCAATCCACGCGCTCGCGAAGAGCGCGACATCTGTTCGCGGCTATCG
2799706 ATCTGGCCAATCCCTGGTTTCAATCCACGCGCTCGCGAAGAGCGCGACGTGTGGTGGCCG
2799766 TTGGCCTGGTGGCGTGGGCATTGTTTCAATCCACGCGCTCGCGAAGAGCGCGACCACATT
2799826 AACCAGAGAAAACTTGGTGATGTCACAGGGTTTCAATCCACGCGCTCGCGAAGAGCGCGA
2799886 CGTTCAGCGTGTATTGATTCCTGCACTGCAAGCGGGTTTCAATCCACGCGCTCGCGAAGA
2799946 GCGCGACACATGTACCCAAAGAGGCTACCTATTTGCTGGATGTTTCAATCCACGCGCTCG
2800006 CGAAGAGCGCGACGCGTATCCATAACACGGCCACTGGGCACAAGAATGTTTCAATCCACG
2800066 CGCTCGCGAAGAGCGCGACCCGGCTGGATGACATTACGCATTTGATTCATGGTGTTTCAA
2800126 TCCACGCGCTCGCGAAGAGCGCGACATTTATGCGCGTATGGTGTGGGGGCTCGACAAGTG
2800186 TTTCAATCCACGCGCTCGCGAAGAGCGCGACTGGTGTTCAGGGATTCGCAGCTAAAGAGA
2800246 GGCAAGTTTCAATCCACGCGCTCGCGAAGAGCGCGACTTTCTTATCCCTTGGTGTCAATT
2800306 TCAATATTAATGTTTCAATCCACGCGCTCGCGAAGAGCGCGACTTACAAGCGGATCTTTG
2800366 TTGTTGTTGAAATACAGTTTCAATCCACGCGCTCGCGAAGAGCGCGACCTGGTAGAGACT
```

Donald Ta
BNFO301

```
2800426 GCAGCCGCAGAAACCCTGTGCGGTTTCAATCCACGCGCTCGCGAAGAGCGCGACCAGGTG
2800486 TCGGGCTCAGAGGGTGATATAGGGTCACTGTTTCAATCCACGCGCTCGCGAAGAGCGCGA
2800546 CCAAGGGGATCGATATTAGTGTTACCGATGCAGTGTTTCAATCCACGCGCTCGCGAAGAG
2800606 CGCGACCCACCAAGATCGCTTAAAGATTTTTTGGCGTTAGTTTCAATCCACGCGCTCGCG
2800666 AAGAGCGCGACCTTTAGGTTTGCCCTAACTCCAAGCCTTTCTGAGTTTCAATCCACGCGC
2800726 TCGCGAAGAGCGCGACTTCGCCCGCTTCGGTCGCTTTGCCGATAATCGCATGTTTCAATC
2800786 CACGCGCTCGCGAAGAGCGCGACTCGGAATGCTTGGTATTGCTGCGGTCGCAGAACTTGT
2800846 TTCAATCCACGCGCTCGCGAAGAGCGCGACACGGCTTCAGCTATTGACTCGTCACCGGTG
2800906 CGGCTGTTTCAATCCACGCGCTCGCGAAGAGCGCGACCTGAAACATACCTTCGCGCCAAG
2800966 CAAATGCTTTTGTTTCAATCCACGCGCTCGCGAAGAGCGCGACCAGATCGGACGCGCTAC
2801026 CACTTCTTGTGCACGATGTTTCAATCCACGCGCTCGCGAAGAGCGCGACCTTCCGTGTTA
2801086 CCTGTCTGGTTGTAGTACCCGCTGTTTCAATCCACGCGCTCGCGAAGAGCGCGACAGGAG
2801146 CTCTTGCCACAGTTGCACACCGGTGGCTTGTTTCAATCCACGCGCTCGCGAAGAGCGCGA
2801206 CTTCTCTTGAAGTTAAAGACGCTGCTACAGATGCAGTTTCAATCCACGCGCTCGCGAAGA
2801266 GCGCGACCGAACAATGCGATACTTGTTGTTCTGTTAACATTGTTTCAATCCACGCGCTCG
2801326 CGAAGAGCGCGACTAAGGCAAACAGCAGAAATTGATGGCCATAGTATTGTTTCAATCCAC
2801386 GCGCTCGCGAAGAGCGCGACACTATCTTGGGCGTGAAGTTGGTCGTCCTGACCTTGTTTC
2801446 AATCCACGCGCTCGCGAAGAGCGCGACTAAATCCGTTAATAGAGCTCTTCTGATGGATAA
2801506 AGTTTCAATCCACGCGCTCGCGAAGAGCGCGACGCGCGGCATTGCAAGGTAGAACTTTTA
2801566 CGATAGCAGTTTCAATCCACGCGCTCGCGAAGAGCGCGACTTTCCCAGGTGGTTAATGTC
2801626 CGTACTTCAGTTTTTGTTTCAATCCACGCGCTCGCGAAGAGCGCGACCCCGCTTTACTGT
2801686 ATAGCCAAATACCAAGAGCGCGTTTCAATCCACGCGCTCGCGAAGAGCGCGACTGGATGA
2801746 ATAGTTACCATCGTACCGCCAGGCGTTGTTTCAATCCACGCGCTCGCGAAGAGCGCGACC
2801806 CACCACCATCTGTTCCTGCCTTGTCCCAATAACGTTTCAATCCACGCGCTCGCGAAGAGC
2801866 GCGACCCGAGATTATCAGCGCACATGCTGAGGCCAGTCTGTTTCAATCCACGCGCTCGCG
2801926 AAGAGCGCGACCATTGACTCTGCGTAGCAGTGATCGCATCCAGGGGTTTCAATCCACGCG
2801986 CTCGCGAAGAGCGCGACTACCAGCGTGTGAGTGGGGCGAACGATATAGCCGAGTTTCAAT
2802046 CCACGCGCTCGCGAAGAGCGCGACACAAACGGGGCTTTGACATACAAAAAACTTGCTGTT
2802106 TCAATCCACGCGCTCGCGAAGAGCGCGACTTGTAGGGCCATCGGCTGTTGGCCCGGCTGG
2802166 CAGTTTCAATCCACGCGCTCGCGAAGAGCGCGACCTTCCGCTGCGCGAATAATCAATTCA
2802226 AATTGGCTGTTTCAATCCACGCGCTCGCGAAGAGCGCGACGTGCTACGAAGACAACCCTC
2802286 ACGATGCTCTTGAGGTTTCAATCCACGCGCTCGCGAAGAGCGCGACTTTCATGTCCGAAT
2802346 TATCAAAAACTTTCTGCGACAGTTTCAATCCACGCGCTCGCGAAGAGCGCGACTGCATTA
2802406 AGGTAACCTATTGGTAAATAAAAAGAAAACTGAAAAATTCCGCTAACCAGGGATATATTT
```

The repeated sequence is highlighted in grey. The repeat goes through an unknown protein highlighted in pink.

The chances of this specific 32 nucleotide sequence occurring in this organism is about 5.42×10^{-20}. So it is actually highly unlikely that it is occurring by chance in such high frequency.

3261290	3261303	GTTTCAATCCACAC
3261358	3261371	GTTTCAATCCACAC
3261424	3261437	GTTTCAATCCACAC
3261491	3261504	GTTTCAATCCACAC
3261557	3261570	GTTTCAATCCACAC
3261624	3261637	GTTTCAATCCACAC
3261690	3261703	GTTTCAATCCACAC
3261758	3261771	GTTTCAATCCACAC
3261825	3261838	GTTTCAATCCACAC
3261891	3261904	GTTTCAATCCACAC
3261960	3261973	GTTTCAATCCACAC
3262028	3262041	GTTTCAATCCACAC
3262094	3262107	GTTTCAATCCACAC
3262161	3262174	GTTTCAATCCACAC
3262228	3262241	GTTTCAATCCACAC
3262294	3262307	GTTTCAATCCACAC
3262360	3262373	GTTTCAATCCACAC
3262428	3262441	GTTTCAATCCACAC
3262494	3262507	GTTTCAATCCACAC
3262562	3262575	GTTTCAATCCACAC
3262628	3262641	GTTTCAATCCACAC
3262695	3262708	GTTTCAATCCACAC
3262761	3262774	GTTTCAATCCACAC
3262829	3262842	GTTTCAATCCACAC
3262895	3262908	GTTTCAATCCACAC
3262960	3262973	GTTTCAATCCACAC
3263026	3263039	GTTTCAATCCACAC
3265050	3265063	GTTTCAATCCACAC
3265118	3265131	GTTTCAATCCACAC
3265185	3265198	GTTTCAATCCACAC
3265252	3265265	GTTTCAATCCACAC
3265319	3265332	GTTTCAATCCACAC
3265387	3265400	GTTTCAATCCACAC
3265454	3265467	GTTTCAATCCACAC
3265520	3265533	GTTTCAATCCACAC
3265588	3265601	GTTTCAATCCACAC
3265655	3265668	GTTTCAATCCACAC
3265721	3265734	GTTTCAATCCACAC
3265787	3265800	GTTTCAATCCACAC
3265854	3265867	GTTTCAATCCACAC
3265922	3265935	GTTTCAATCCACAC
3265989	3266002	GTTTCAATCCACAC
3266056	3266069	GTTTCAATCCACAC
3266122	3266135	GTTTCAATCCACAC
3266188	3266201	GTTTCAATCCACAC
3266255	3266268	GTTTCAATCCACAC
3266321	3266334	GTTTCAATCCACAC

Figure 4: A. *vinelandii* DJ MATCHES-OF-PATTERN output 1

My initial search for any matches of the sequence found in C. *japonicus*

Figure 5: A. *vinelandii* DJ MATCHES-OF-PATTERN output after extending with "*"

3261020	3261055	CTGTTTCAATCCACACGCCCGCATGGGGCGTGACCG
3261088	3261123	GGGTTTCAATCCACACGCCCGCATGGGGCGTGACTG
3261154	3261189	ATGTTTCAATCCACACGCCCGCATGGGGCGTGACAT
3261222	3261257	CCGTTTCAATCCACACGCCCGCATGGGGCGTGACAC
3261288	3261323	CAGTTTCAATCCACACGCCCGCATGGGGCGTGACAT
3261356	3261391	GTGTTTCAATCCACACGCCCGCATGGGGCGTGACAG
3261422	3261457	ATGTTTCAATCCACACGCCCGCATGGGGCGTGACGG
3261489	3261524	GAGTTTCAATCCACACGCCCGCATGGGGCGTGACGC
3261555	3261590	GTGTTTCAATCCACACGCCCGCATGGGGCGTGACCT
3261622	3261657	CCGTTTCAATCCACACGCCCGCATGGGGCGTGACCG
3261688	3261723	AAGTTTCAATCCACACGCCCGCATGGGGCGTGACAG
3261756	3261791	CTGTTTCAATCCACACGCCCGCATGGGGCGTGACGT
3261823	3261858	TTGTTTCAATCCACACGCCCGCATGGGGCGTGACGG
3261889	3261924	GCGTTTCAATCCACACGCCCGCATGGGGCGTGACGA
3261958	3261993	CTGTTTCAATCCACACGCCCGCATGGGGCGTGACGC
3262026	3262061	TGGTTTCAATCCACACGCCCGCATGGGGCGTGACTC
3262092	3262127	TTGTTTCAATCCACACGCCCGCATGGGGCGTGACCC
3262159	3262194	ATGTTTCAATCCACACGCCCGCATGGGGCGTGACTG
3262226	3262261	GCGTTTCAATCCACACGCCCGCATGGGGCGTGACTG
3262292	3262327	GTGTTTCAATCCACACGCCCGCATGGGGCGTGACGC
3262358	3262393	ACGTTTCAATCCACACGCCCGCATGGGGCGTGACTA
3262426	3262461	ACGTTTCAATCCACACGCCCGCATGGGGCGTGACAG
3262492	3262527	CAGTTTCAATCCACACGCCCGCATGGGGCGTGACCA
3262560	3262595	ACGTTTCAATCCACACGCCCGCATGGGGCGTGACTC
3262626	3262661	CCGTTTCAATCCACACGCCCGCATGGGGCGTGACGA
3262693	3262728	GAGTTTCAATCCACACGCCCGCATGGGGCGTGACCT
3262759	3262794	AGGTTTCAATCCACACGCCCGCATGGGGCGTGACGC
3262827	3262862	GCGTTTCAATCCACACGCCCGCATGGGGCGTGACCC
3262893	3262928	ACGTTTCAATCCACACGCCCGCATGGGGCGTGACCA
3262958	3262993	CAGTTTCAATCCACACGCCCGCATGGGGCGTGACAT
3265048	3265083	ATGTTTCAATCCACACGCCCGCATGGGGCGTGACCA
3265116	3265151	CGGTTTCAATCCACACGCCCGCATGGGGCGTGACTT
3265183	3265218	GTGTTTCAATCCACACGCCCGCATGGGGCGTGACCA
3265250	3265285	CCGTTTCAATCCACACGCCCGCATGGGGCGTGACGT
3265317	3265352	CTGTTTCAATCCACACGCCCGCATGGGGCGTGACCT
3265385	3265420	AAGTTTCAATCCACACGCCCGCATGGGGCGTGACCG
3265452	3265487	CCGTTTCAATCCACACGCCCGCATGGGGCGTGACTG
3265518	3265553	GTGTTTCAATCCACACGCCCGCATGGGGCGTGACAT
3265586	3265621	GCGTTTCAATCCACACGCCCGCATGGGGCGTGACTC
3265653	3265688	TTGTTTCAATCCACACGCCCGCATGGGGCGTGACGC
3265719	3265754	AAGTTTCAATCCACACGCCCGCATGGGGCGTGACCT
3265785	3265820	GGGTTTCAATCCACACGCCCGCATGGGGCGTGACAA
3265852	3265887	AAGTTTCAATCCACACGCCCGCATGGGGCGTGACGT
3265920	3265955	TTGTTTCAATCCACACGCCCGCATGGGGCGTGACCC
3265987	3266022	ATGTTTCAATCCACACGCCCGCATGGGGCGTGACTC
3266054	3266089	CTGTTTCAATCCACACGCCCGCATGGGGCGTGACTG
3266120	3266155	AGGTTTCAATCCACACGCCCGCATGGGGCGTGACGC
3266186	3266221	TTGTTTCAATCCACACGCCCGCATGGGGCGTGACAC
3266253	3266288	GTGTTTCAATCCACACGCCCGCATGGGGCGTGACCG
3266319	3266354	AAGTTTCAATCCACACGCCCGCATGGGGCGTGACCC

You can see in the figure to the left that the repeat highlighted in yellow is very similar to that found in C. *japonicus*

The chances of this specific 32 nucleotide sequence occurring by chance is about 5.15×10^{-19}. So it is highly unlikely that this repeat occurred by chance. This is different from that of C. *japonicus* because the ratios of G, C, T, and A are different. This organism has a higher abundance of G and C compared to T and A.

Figure 6: Pseudomonas stutzeri Matches of Pattern output

4056652	4056687	TGGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAA	F
4056720	4056755	ACGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAA	F
4056786	4056821	CGGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAA	F

4056853	4056888	AAGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAC	F
4056923	4056958	GCGTCGCGCCCCACGCGGGCGCGTGGATTGAAACCA	F
4056989	4057024	CCGTCGCGCCCCACGCGGGCGCGTGGATTGAAACTG	F
4057055	4057090	ACGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAC	F
4057121	4057156	TTGTCGCGCCCCACGCGGGCGCGTGGATTGAAACCT	F
4057187	4057222	CGGTCGCGCCCCACGCGGGCGCGTGGATTGAAACTG	F
4057254	4057289	ATGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAA	F
4057320	4057355	GAGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAA	F
4057386	4057421	TCGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAA	F
4057452	4057487	TGGTCGCGCCCCACGCGGGCGCGTGGATTGAAACTT	F
4057518	4057553	GCGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAC	F
4057585	4057620	TGGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAC	F
4057651	4057686	GGGTCGCGCCCCACGCGGGCGCGTGGATTGAAACTT	F
4057717	4057752	GAGTCGCGCCCCACGCGGGCGCGTGGATTGAAACTC	F
4057786	4057821	CTGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAT	F
4057852	4057887	CGGTCGCGCCCCACGCGGGCGCGTGGATTGAAACTG	F
4057918	4057953	GAGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAT	F
4057983	4058018	ACGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAG	F
4058049	4058084	CCGTCGCGCCCCACGCGGGCGCGTGGATTGAAACTT	F
4058115	4058150	GCGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAG	F
4058180	4058215	TCGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAT	F
4058247	4058282	CCGTCGCGCCCCACGCGGGCGCGTGGATTGAAACTG	F
4058314	4058349	CGGTCGCGCCCCACGCGGGCGCGTGGATTGAAACTG	F
4058381	4058416	GTGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAT	F
4058450	4058485	CTGTCGCGCCCCACGCGGGCGCGTGGATTGAAACCA	F
4058517	4058552	TAGTCGCGCCCCACGCGGGCGCGTGGATTGAAACGA	F
4058584	4058619	CAGTCGCGCCCCACGCGGGCGCGTGGATTGAAACCT	F
4058651	4058686	CTGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAT	F
4058720	4058755	ATGTCGCGCCCCACGCGGGCGCGTGGATTGAAACCT	F
4058786	4058821	GGGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAA	F
4058853	4058888	AGGTCGCGCCCCACGCGGGCGCGTGGATTGAAACAG	F
4058918	4058953	CCGTCGCGCCCCACGCGGGCGCGTGGATTGAAACCT	F
4058983	4059018	GCGTCGCGCCCCACGCGGGCGCGTGGATTGAAACGT	F
4059049	4059084	AAGTCGCGCCCCACGCGGGCGCGTGGATTGAAACGA	F
4059114	4059149	TGGTCGCGCCCCACGCGGGCGCGTGGATTGAAACTA	F
4059247	4059282	ACGTCGCGCCCCACGCGGGCGCGTGGATTGAAACTT	F
4059313	4059348	GAGTCGCGCCCCACGCGGGCGCGTGGATTGAAACGG	F

The repeated sequence is highlighted in yellow to the right. You can see that it is the exact compliment to the repeated sequence shown in Figure 3. This is an interesting find because these organisms are different, but they are just in the same group.

The chances of this happening completely by chance is 5.42×10^{-20} so it is highly unlikely this sequence occurred by chance.

Figure 7: Acinetobacter Count of K-mers to determine most repeated sequences output

```
Warning: Threshold set to 3 to save memory
COUNT           WORD
   91   ATCGCATAGATGATTTAGAA
   91   CATCGCATAGATGATTTAGA
   91   CGTCATCGCATAGATGATTT
   91   GTCATCGCATAGATGATTTA
   91   TCATCGCATAGATGATTTAG
   91   TCGCATAGATGATTTAGAAA
   91   TCGTCATCGCATAGATGATT
   91   TTCGTCATCGCATAGATGAT
   90   GTTCGTCATCGCATAGATGA
```

Figure 8: Acinetobacter Output of the Repeats found in this organism

```
2447995 ACATATGTAGTAAAATCAAATAAAATCAAACAATTAACTCAAGTGATTCATAACGAAGTA
2448055 TTTTTACTCATTAAAAGCTTATATAATTGATATCAAGGGTTTTGTTTTGACTTAACTCTA
2448115 GTTCGTCATCGCATAGATGATTTAGAAAGACACCAAAGGTAATAAAGCTATGAAAGAATA
2448175 GTTCGTCATCGCATAGATGATTTAGAAATTTACTCTTATTATACTATTACCCCTAACCCC
2448235 GTTCGTCATCGCATAGATGATTTAGAAATCCAGCTAAAATCGTTTGAGGGTGAAACTCCT
2448295 GTTCGTCATCGCATAGATGATTTAGAAAATGATTTCGAAAGGCTCTCCGAGTACGTTATT
2448355 GTTCGTCATCGCATAGATGATTTAGAAAATTCCCAGCATTCACGCTGAGTGCTTCGGCAC
2448415 GTTCGTCATCGCATAGATGATTTAGAAATGTGCAGCCGTTTGGCGCGCCCCAGATATGCG
2448475 GTTCGTCATCGCATAGATGATTTAGAAAAGGAACCGTGGCAGATTGCGTTAATATGTTAG
2448535 GTTCGTCATCGCATAGATGATTTAGAAATAACGATGGAATAACGTTCAAAGAATCTAACG
2448595 GTTCGTCATCGCATAGATGATTTAGAAAAATTCATGAAAGATCATTCGCTGTGTTTGGGG
2448655 GTTCGTCATCGCATAGATGATTTAGAAAATTTGCCGCTTTGAATATTTGATGCACCTGCT
2448715 GTTCGTCATCGCATAGATGATTTAGAAAAAATCGATGAGGGACAACATCAGGCACTCGAC
2448775 GTTCGTCATCGCATAGATGATTTAGAAAACAGGGCAGGGAAATAACCAAAAATCGATATA
2448835 GTTCGTCATCGCATAGATGATTTAGAAAGACATAGGAACGATATGAAGATGATTTTTTTT
2448895 GTTCGTCATCGCATAGATGATTTAGAAAATCAAGCTATCGTCATTTGGCCGATACACAGC
2448955 GTTCGTCATCGCATAGATGATTTAGAAATCTGCCATGCATACAATTTGATTTGGCTGCGT
2449015 GTTCGTCATCGCATAGATGATTTAGAAAAATCATCAATATCTTTTTGCGCTTTGCGTGAA
2449075 GTTCGTCATCGCATAGATGATTTAGAAATCTCACGTACAAAAAAAAATCCTATTTGATGT
2449135 GTTCGTCATCGCATAGATGATTTAGAAAGCGATTGAATACCGATAGATCGGGGATATTAA
2449195 GTTCGTCATCGCATAGATGATTTAGAAAATACACTACATTGAACTGCTCGGACTTAAGCA
2449255 TGTTCGTCATCGCATAGATGATTTAGAAAAAAAAAAGTGTAGCCAACTTCATACAGTTAC
2449315 CGTTCGTCATCGCATAGATGATTTAGAAACAGGTGGCAGCGTTCCATTTTCGGGGGCAAA
2449375 TGTTCGTCATCGCATAGATGATTTAGAAAAAAACCACATTATAAGGCTCGGTAAATGTGT
2449435 AGTTCGTCATCGCATAGATGATTTAGAAAATGAAAATAAGCCCCAATATTGTCAGTGTTC
2449495 CGTTCGTCATCGCATAGATGATTTAGAAAGTTTCCGCGTCATTCGGGTACAGTTGCGACA
2449555 TGTTCGTCATCGCATAGATGATTTAGAAATTGAAACCTATGAACTTTGTGTTATACGTGT
2449615 CGTTCGTCATCGCATAGATGATTTAGAAACTTATCAAAATCGGTGGGATCTTTGTCGTAC
2449675 TGTTCGTCATCGCATAGATGATTTAGAAAGAATTATGCTTTAAAAAATCCTTTCGCGGGT
2449735 AGTTCGTCATCGCATAGATGATTTAGAAAAATCCGATTTCTGCTGTTGCTGGGGTTAGAG
2449795 CGTTCGTCATCGCATAGATGATTTAGAAAATGTACTATAAGTCACATGGTAAAGACACGA
2449855 AGTTCGTCATCGCATAGATGATTTAGAAAGAAACGTTGAATCCAGAACCAGCAATCCCAG
2449915 CGTTCGTCATCGCATAGATGATTTAGAAAAAACTGTGGAGCATTACATCTACCATACTGC
2449975 CGTTCGTCATCGCATAGATGATTTAGAAATAAAACAGTCAATGTTAATTGGGGTGAACAA
2450035 TGTTCGTCATCGCATAGATGATTTAGAAAGCGGTAGCTGGCGCGGTGTTTGCGTTTTTTG
2450095 GGTTCGTCATCGCATAGATGATTTAGAAATATAACTAGCATGTCAGAAATAAAACTATCC
2450155 GGTTCGTCATCGCATAGATGATTTAGAAACGTTGGTACTGTTGCAGGTGGTGCATTGGGG
2450215 AGTTCGTCATCGCATAGATGATTTAGAAAGACTCCGCTACTTAAGAAAGAGAGCATAGGT
2450275 GGTTCGTCATCGCATAGATGATTTAGAAATAGAAGTAACTTACGATAACATCTTTGGCGC
2450335 CGTTCGTCATCGCATAGATGATTTAGAAATCAAGCATGTGATCACTAATGATTCGGTTTT
2450395 TGTTCGTCATCGCATAGATGATTTAGAAATATACTCCTTATATGTAATTTACGCGTAAAC
2450455 CGTTCGTCATCGCATAGATGATTTAGAAACACTACATTTATACCCGCCGTTTACGCTCTT
2450515 AGTTCGTCATCGCATAGATGATTTAGAAAGTTTAATGTGGCGTTCAGGTCTTGTTCGCCA
2450575 AGTTCGTCATCGCATAGATGATTTAGAAAACTCAGTTGACCAATCTTACTGCTTCACTTA
2450635 AGTTCGTCATCGCATAGATGATTTAGAAAAGAAGATTTGGTGGGCAAAAATATGGAATAT
2450695 AGTTCGTCATCGCATAGATGATTTAGAAAATTCTTAGCTGCATCACGCAAGATTTGCTTT
2450755 TGTTCGTCATCGCATAGATGATTTAGAAACTCATCGAAACATACATTGAGAAAAATCATT
2450815 TGTTCGTCATCGCATAGATGATTTAGAAAAATCATCATCGACCGCAGTATTGAAGCGAAG
2450875 CGTTCGTCATCGCATAGATGATTTAGAAAAGCCCTTCGTATATTTGAATAGTGCATTGGC
2450935 TGTTCGTCATCGCATAGATGATTTAGAAAAAAAATACCCGCGCCCAAGTGATCCTGAAGA
2450995 GTTCGTCATCGCATAGATGATTTAGAAAACCATATAGAATTGTTAACTTTTGTAAATAA
2451055 GTTCGTCATCGCATAGATGATTTAGAAAGATCAAAACAACAAGCGTACCAATGATGCCGA
2451115 GTTCGTCATCGCATAGATGATTTAGAAAACAAGGGATGTATTGACCAGGTGTGAGCGCAA
2451175 GTTCGTCATCGCATAGATGATTTAGAAAATTCTTGAGCCGCCTGCAGATTTGTTATGTCA
2451235 GTTCGTCATCGCATAGATGATTTAGAAAATGGTTCGGGGTTGTAGCTGTACGCCCCAGAT
2451295 GTTCGTCATCGCATAGATGATTTAGAAAAAGAGCAAAAGGTAACTTGGATCTACCGCCAC
```

Donald Ta
BNFO301

```
2451355 GTTCGTCATCGCATAGATGATTTAGAAACACGGAAATTGGAATGATGATTTCGACGGTAA
2451415 GTTCGTCATCGCATAGATGATTTAGAAATTGTTGAGCAGCAGAACGGCCTTTTACCAACC
2451475 GTTCGTCATCGCATAGATGATTTAGAAAAGATACCTCAGTCCAAGCTGCTGAATTTTATC
2451535 GTTCGTCATCGCATAGATGATTTAGAAAAAGAGACAACAGGGCTTATTAAAGTAACTTGT
2451595 GTTCGTCATCGCATAGATGATTTAGAAAAAGTTTTATTTAAGCCCAAAGCTAAAGATAGT
2451655 GTTCGTCATCGCATAGATGATTTAGAAAGTTAGCTGCACAAGCTCTGGGACTTTAATAAA
2451715 GTTCGTCATCGCATAGATGATTTAGAAAAATCGCTAACCAGTAGAACCCGCGTAGCAGCG
2451775 GTTCGTCATCGCATAGATGATTTAGAAAAAGCGTTGCGAGCGCTCAAAAAGTGGCTGATC
2451835 GTTCGTCATCGCATAGATGATTTAGAAAGTCTACCAAAGCGAAAGTATCATTTTCAATGA
2451895 GTTCGTCATCGCATAGATGATTTAGAAATGTATCGGAGCTACGTCAGAAGGTCAAGCACA
2451955 GTTCGTCATCGCATAGATGATTTAGAAAAGGTCGATTTATCATAAACATCGGGCACGATA
2452015 GTTCGTCATCGCATAGATGATTTAGAAAGCCAGAAATTTTGACACTTGCGTTTAGCAATA
2452075 GTTCGTCATCGCATAGATGATTTAGAAAAGATTGTCTCTAAATTTAACGCGTGGCTTTGT
2452135 GTTCGTCATCGCATAGATGATTTAGAAAAAAGCCGAGCCCAACTTTTGACGCACAAAAAG
2452195 GTTCGTCATCGCATAGATGATTTAGAAAGTCAGTGATTGCTTTCATTGCCGTAGCTACGT
2452255 GTTCGTCATCGCATAGATGATTTAGAAAATCCGCGCCCAATTTGTCCCACCAATCTTTTT
2452315 GTTCGTCATCGCATAGATGATTTAGAAAGATTCCATAGAACGTACCATTGACGCGCAACA
2452375 GTTCGTCATCGCATAGATGATTTAGAAATGGATCTCTGCAGAAATCACATTGTCCAAATA
2452435 GTTCGTCATCGCATAGATGATTTAGAAAAACAGGCGTTACTGAGCTATGTGTCGTTAAAA
2452495 GTTCGTCATCGCATAGATGATTTAGAAAAAGCATGCCTTGATGCATACAACAAAATTGCC
2452555 GTTCGTCATCGCATAGATGATTTAGAAATGCGAGTTCAAACTTCTTTAAAGATGCAACAT
2452615 GTTCGTCATCGCATAGATGATTTAGAAACGTGGAATCATAATCATAAGCTTCACCGACAC
2452675 GTTCGTCATCGCATAGATGATTTAGAAAGATCAGTGGCGCGTCTACAGTGAGCGAGTGGG
2452735 GTTCGTCATCGCATAGATGATTTAGAAAATAATTGCAACAACAGCATAATATACATACCA
2452795 GTTCGTCATCGCATAGATGATTTAGAAACTTACTTTCGCTTGCGCTTCGTTACGAATGCC
2452855 GTTCGTCATCGCATAGATGATTTAGAAATCAACCAGGATCGGATAACCATCAATTCTAAA
2452915 GTTCGTCATCGCATAGATGATTTAGAAAAACAGGCGTTACTGAGCTATGTGTCGTTAAAA
2452975 GTTCGTCATCGCATAGATGATTTAGAAAAAGCATGCCTTGATGCATACAACAAAATTGCC
2453035 GTTCGTCATCGCATAGATGATTTAGAAACAAATGTAATCAGGATTAGTCGATTGCAGCGT
2453095 GTTCGTCATCGCATAGATGATTTAGAAAAGATCGCCTGTGCGTAGGTCAACTGCACCATT
2453155 GTTCGTCATCGCATAGATGATTTAGAAAAGCTGAACACGCCGTTTTTTAACTTCCGCCAT
2453215 GTTCGTCATCGCATAGATGATTTAGAAAATGCACCTGATCCTGCCCAATGAGGGATTTAC
2453275 ATTCGTCATCGCATAGATGATTTAGAAATGATGGTGCAGGAACCACAGCAACATCAGTCA
2453335 GTTCGTCATCGCATAGATGATTTAGAAAGATTGAAATACTATTAAGGCTGTTCGTAAAGC
2453395 GTTCGTCATCGCATAGATGATTTAGAAAACACACGCTGCCAATTCTTCGTTAGAGTGTAT
2453455 GTTCGTCATCGCATAGATGATTTAGAAAAGCAGTAAAAGCCATGACCGTTAAGATCGCTC
2453515 GTTCGTCATCGCATAGATGATTTAGAAATATTAAAAGCAACATCGATAAGATCTAGCTGT
```

You can see that the repeat is highlighted in grey. This repeat in particular seems to be mostly intragenic (found within genes) as there actually seems to be little extragenic space.

This repeat also seems to only be 28 nucleotides in length as opposed to 32 like in the other organisms examined.

Works Cited

[1]Aranda-Olmedo, I., Tobes, R., Manzanera, M., Ramos, J. L., & Marqués, S. (2002). Species-specific repetitive extragenic palindromic (REP) sequences in Pseudomonas putida. *Nucleic Acids Research*, *30*(8), 1826–33. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=113213&tool=pmcentrez&rendertype=abstract

[2]Petrillo, M., Silvestro, G., Di Nocera, P. P., Boccia, A., & Paolella, G. (2006). Stem-loop structures in prokaryotic genomes. *BMC Genomics*, *7*, 170. doi:10.1186/1471-2164-7-170

[3]Sorek, R., Kunin, V., & Hugenholtz, P. (2008). CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea, *6*(march), 181–186.

[4]BioBike PhAnToMe http://biobike.csbc.vcu.edu/

[5] Basic Local Alignment Search Tool. http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome