Rough Draft

Consensus Short Dispersed Repeats Sequences in E. Coli near Important Genes

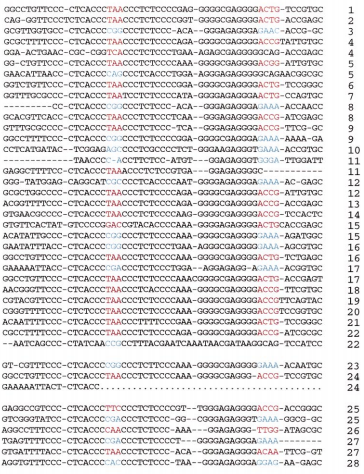
**Introduction**

Various kinds of sequences are found in DNA; both eukaryotic and prokaryotic. Some of these sequences includes many forms of nucleotides that code for various genes and nucleotides that helps the genome stability of the organism. However, there are certain kinds of sequences that are shown to be “repeated” across an organism’s DNA. These are known as “Repeated Sequences” and they come in many forms, such as: Tandem Repeats, Short Dispersed Repeats, Cluster Repeats, etc. It was believed that repeats were only found in Eukaryotic genomes due to prokaryotes having a small nucleotide (Ying Xu et al.). However, it appears in prokaryotic genomes as well as stated by some articles. It may be interesting to see repeats in prokaryotes as there are labeled for genomic stability (Odahara, Masaki et al.). In doing so, there is some intent to discover these repeated sequences, mainly for this paper, short dispersed repeats.

Short Dispersed Repeats (SDR) are repeats of a DNA sequence with gaps in between them. Looking at FIG 1, it shows the sequence of CGGAGAGG in pink located across one strand of the black colored DNA sequence. It is noticeable to see that it appears twice in the strand with a nucleotide gap in between them. That is what short dispersed repeats are as they are dispersed from one another.

**FIG 1: Model of SDR (Sequence Assembly – Wikipedia)**

For the majority of the paper, E. coli SDR sequences will be looked at. The reason being is that there have been debate in deciding if there are any major SDR sequences in E. coli. One article by Belkum states that after looking over the genome of E. coli K-12, only 5 repeats were found per loci of a gene (Alex van Belkum et al.) whereas other papers state that 28-bp SDR sequences were found where 14 repeats where found per loci where there were 32 bp gaps (Bachellier, Sophie et al). These different sequences can be seen on FIG 2. Ying Xu also pointed out highly overrepresented repeat sequences that took the form of Highly Iterated Palindrome (HIP). These HIP had consensus sequences of GCGATCGC or GGCGATCGCC for many cyanobacteria and having GCTGGTGG for E. coli. After looking for these SDR sequences, Ying Xu et al. described SDR sequences of having “significant roles” for an organism’s genome. It seems that the distribution of dispersed repeats may lead to specific functional roles of the repeats. They may be related to specific cellular processes and/or they have a role in transcription as multiple G and C bases may lead to changes in binding for organisms. The most important factor is its function in recombination as iterated palindromic sequences can be broken apart at multiple locations but can be fixed through homologous recombination. Due to these hypothetical events, it seems important to see the placement of SDR sequences in E. coli and see what their consensus SDR sequences is and where are they located relative to a gene.

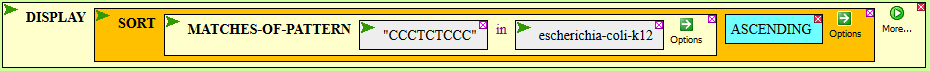


**FIG 2: 28 SDR sequences found in E. coli with mismatches (Bachellier, Sophie et al)**

**Methods and Experiment**

The main forms of looking at repeats of the E. coli bacteria is through biobike and utilizing the SEQUENCE SIMILAR TO, COUNT OF, MATCHES OF PATTERN functions. For the experiment, the 4 E. coli genomes that were looked at were: k-12, f11, hs, and ATCC-8739. The sequences were compared to one another with k-12 being the primary sequence to observe. By using the sequences of FIG 2, a common sequence of CCCTCTCCC will be used. Counts of K-mers will be used to see the amount of 7, 8, or 9 bp lengths were found in the organisms as these counts seemed to be the standardized lengths found in literature. Mismatches may be taken into consideration as determined from FIG 2.

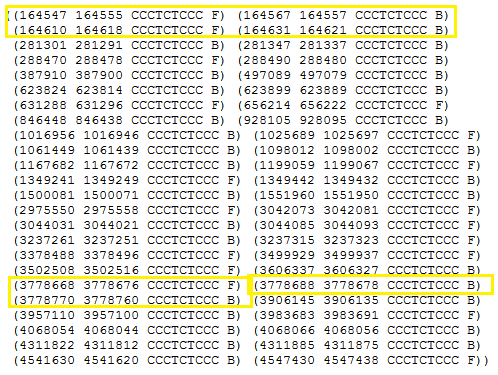
Here is the general function that was used to search for the sequences:



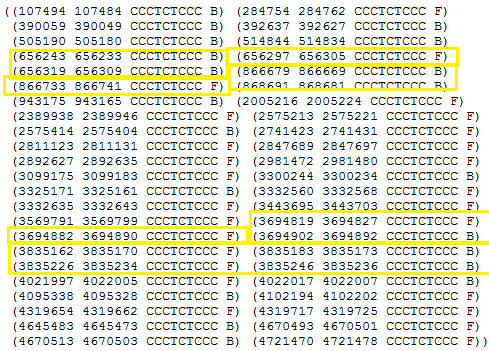
**Results and Discussion**

Through using COUNT OF function, 48 CCCTCTCCC sites were found in K-12, 46 in ATCC, 17 in F11, and 21 in HS. Besides ATCC, all other genomes had similar amounts of the reported SDR sequence. Besides HS, all other genome had similar amounts of reported SDR sequence.

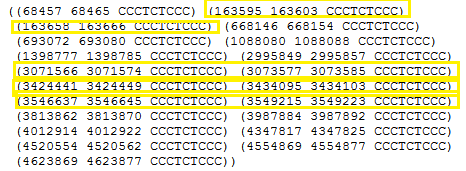
K-12:



ATCC:

****

HS:

****

**FIG 3: CCCTCTCCC consensus sites for E. coli K-12, ATCC, and HS.**

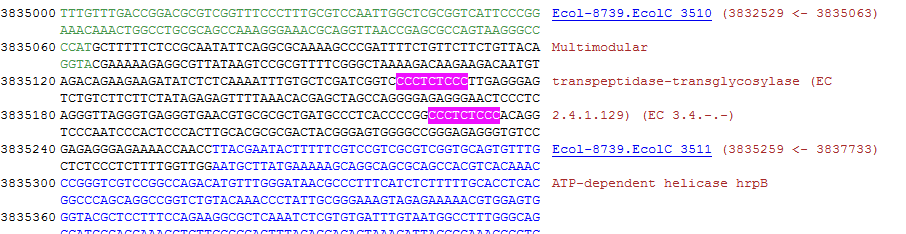
F11 was not included since it did not have any sequence-coordinate matches of interest. E. coli K-12, ATCC, and HS had more promising results. From the results given by the biobike programs, short dispersed sequences are given by coordinate who appear close approximation to one another. All of the highlighted boxes shows sequences that appear near one another in sets of 3 or higher. There are mainly matches of 2 but the best match shows the one with 4 on K-12 of the 164000 bp positions and another 4 sets one on ATCC of the 3835000 positions. HS had mainly sets of 2.

When looking at these positions, SEQUENCE OF function was used for K-12 and ATCC to see if there is a specific gene around those repeats.

K-12:



ATCC:



**FIG 4: Sequence sites of the hypothetical SDR for K-12 and ATCC**

Results show that hypothetical SDR sequences show up near Multimodular transpeptidase-transglycosylase. However, in K-12, the sequences appears before the gene and in ATCC, the sequences appear after the gene. Another thing to look at is that 2 additional CCCTCTCCC sequences are not observed in the ATCC sequence. It was determined to be there before in FIG 3. However, there does seem to be a correlation between the hypothetical SDR sequences seen by Bachellier, Sophie et al. It appears near the multimodular gene whose function I am not quite knowledgeable with. It might have been important if all of the E. coli strains were checked with this sequence; even HS and F11. This sequence might have an important to specific genes as stated by Kolvhanov, Odahara, and Ying Xu.

My results also confirm with Belkum in which only a set of 4 SDR sequences were obtained near a locus of a gene. However, using mismatches as used by Bachellier might have proven useful as it might have shown the 28 sequences located in their sample. Further testing may be done by performing more/different functions in biobike that may help identify SDR sites readily. The results of this experiment do hint towards that SDR sequences may have roles with specific genes. However, determining if it is recombination, transcription, etc. will be another task to be done in the future.

References

1. Bachellier, Sophie et al. (1999). “Short palindromic repetitive DNA elements in enterobacteria: a survey.” Res. Microbiol.

<http://ac.els-cdn.com/S092325089900128X/1-s2.0-S092325089900128X-main.pdf?_tid=40cd8204-ed74-11e4-a71b-00000aab0f6b&acdnat=1430204608_34ab1825a3102cb9c30f957f13913bf6>

1. Alex van Belkum et al. (1998). “Short-Sequence DNA Repeats in Prokaryotic Genomes.”

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC98915/>

This article is basically a background check article. I learn about the uses and functions of short DNA repeats in genomes for bacteria.

1. Kolvhanov et al. “Computer Analysis of Genetic Macromolecules: Structure, Function, and Evolution.

<https://books.google.com/books?id=crip5tRcF0YC&pg=PA332&lpg=PA332&dq=Dispersed+Repeats+prokaryotic&source=bl&ots=03DCQ7xPt4&sig=1gNGaJ3mY45Jng_SDtCt4KZm9Eo&hl=en&sa=X&ei=d08dVeudDYOMNpSggLAE&ved=0CDQQ6AEwAw#v=onepage&q=Dispersed%20Repeats%20prokaryotic&f=false>

This is another article that explains more in detail on short sequence repeats, more specifically, short dispersed repeats, as most articles I ran into talked mainly about tandem repeats.

1. Mazon, Gerard et al. (2012). “The Rad1-Rad10 nuclease promotes chromosome translocations between dispersed repeats.”

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3443319/>

This was the article that gave me the idea about recombination happening between dispersed repeats. I wondered to myself, “what would happen? Is there a major thing that will happen to a bacteria if recombination occurs? Or is there a functional significance of recombination through these repeats?

1. Odahara, Masaki et al. (2015). “RECG Maintains Plastid and Mitochondrial Genome Stability by Suppressing Extensive Recombination between Short Dispersed Repeats.”

<http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005080>

This is the main article I am looking at to see the experiment in which the researchers performed recombination between SDR and seeing what happens to the Ecoli genome.

1. “The Origin and Possible Functional Role of Short Dispersed Repeats.”

<http://www.vcu.edu/csbc/bbsi/people/students/0506/m_kato/BBSI%20final%20report.pdf>

While hunting for some papers, I saw this on the internet. It is report on SDR sequences while using biobike to see what the functions of SDR are. I even see your name in the references.

1. Ying Xu et al. “Computational Methods for Understanding Bacterial and Archaeal Genomes.”

<https://books.google.com/books?id=GH99OB14Q3UC&pg=PA29&lpg=PA29&dq=short+dispersed+repeats+bacteria&source=bl&ots=WxmIT9DTpT&sig=BFiOwgBKj3-6IZiphzK2LNjUJTw&hl=en&sa=X&ei=m6QaVZGXH4uYNvOzgJgC&ved=0CDwQ6AEwBA#v=onepage&q=short%20dispersed%20repeats%20bacteria&f=false>

Another background paper; however, it goes over computational methods in searching SDR sequences and functional assessments.