* CRISPR Spacer Identification Project
* Nico Crudele
* Computation in the Sciences
* Tiffany Timbers
* Abstract:
* *The CRISPR adaptive immunity has become a major player in the microbiology community as a genome-engineering tool. CRISPR uses a guided endonuclease to insert phage DNA aka ‘spacers’ between short tandem repeats, of fixed lengths. Here, the idea of creating a program capable of finding spacers is explored, through regular expressions in python. Possible new CRISPR software is discussed, along with the implications of increasing the throughput of CRISPR genome scanning and identification.*

Introduction:

* The CRISPR adaptive immune system has become very popular worldwide due to it extremely affordable and efficient nature as a genome-engineering tool. CRISPR is an acronym for Clustered Regularly Interspaced Palindromic Repeats. In Bacteria, this system was found operating naturally as an acquired immune system2. The CRISPR immune system places short pieces of phage DNA in between short repeats of DNA. When we look at the CRISPR array in the genome, we see the pattern follows: CRISPR repeat, spacer, CRISPR repeat, spacer, etc. Here I hypothesize that I can create a code to identify the genomic location of CRISPR repeats, and show me the 1st spacer that appears between the 1st and 2nd repeats.
* Methods:
* I was quite unsure as to how to deal with genomic data, near the beginning of the class, as most of the data we dealt with at that point was numerical, and involved relating numbers in different ways. Fortunately, after learning about regular expressions, I found that they would be useful for relating strings of DNA sequences, which do not represent numerical value, but instead represent the nitrogenous bases Adenine, Thymine, Cytosine and Guanine, as the letters A,T,C,G.

After some research, I came across a program online, created by the University of Paris called CRISPRdb3, which is supposed to show all CRISPR repeats for each genome in the NCBI genome database, and the respective spacers found between repeats.

After much trial and error, and hours on Stack-Overflow, I ended up with a relatively simple, straight-forward code, that would print me the beginning nucleotide location, end nucleotide location, and sequence of the spacer between the 1st and 2nd CRISPR repeats. First, the data is munged, to remove line indentations, which are present at a fixed character length per line in Fasta format files, which are a standard format for storing genome information. This is followed by a series of code for finding the beginning location of the 1st and 2nd CRISPR repeats. Unfortunately, this code would throw a complex error message when asked to print past the second iteration of the repeat. With this hinderance, the code is only capable of finding the 1st spacer in that particular CRISPR array. This part of the code provides two variables, which store the beginning location of both iterations of the CRISPR repeat. Then, a similar piece of code creates two more variables, which store the end location of both recorded repeats. A variable called ‘spacer’ is then created, which is the whole genome sequence, from the end point of the 1st repeat, the start point of the 2nd repeat. As the space between two repeats, this is definitively referred to as a spacer. A plot was created, which shows the beginning and end locations of the spacer, on two different bars. This graph shows relatively how far in the genome the specified CRISPR and spacer are, which is neat, but not the most important part of the project.

Results:

The most interesting finding of the project was that my program would find the same CRISPR repeat sequences as CRISPRdb, however the spacers between those repeats would be different, each time that I tested. Separately, my script was able to find all sequences listed as repeats on CRISPRdb, but often would only match that sequence once. If there is only one iteration of the repeat, then it is not a repeat, and that is a bug. This should be further investigated through visual analysis of the whole CRISPR array, i.e. visually parcing out the spacers and repeats, as both these programs find different spacers between the repeats. Separately, for the *Acidithiobacillus ferrooxidans* bacterium, which is one that I used as a test, CRISPRdb listed the 1st and 2nd CRISPR systems in the bacterium as separate, however, after one of my scripts was not operating properly, I found that the 1st and 2nd CRISPR systems used the same repeat, which makes them part of the same system. This is another potential bug in CRISPRdb that should be further investigated.

Future Directions:

Future directions into CRISPR identification code would be to make a modular script that can take in a genome, find repetitive sequences of DNA, scan for Cas genes, which are a fundamental part of CRISPR systems, and any other genomic data that would suggest that the portion of DNA being scanned includes a CRISPR array. This would be useful in cataloging new CRISPR systems in bacterial genomes, as they are sequenced. Already, many variations of CRISPR systems with different endonucleases, other related proteins, and CRISPR sequences have been catalogued, some showing more promise than others in terms of application. The most commonly used endonuclease today is Cas9, which is guided by an RNA strand which must be synthesized inside the operating cell from two other RNA’s, a complex procedure. Recently, CPF1, another endonuclease was discovered operating in bacteria4. CPF1 takes a single RNA guide, which is much simpler to synthesize and use than the duplex necessary for Cas9, which makes experimentation with CRISPR more efficient. Strains of this protein have already been optimized for operation in human cells.

If we can create programs to efficiently scan genomes for characteristics of CRISPR arrays, we’ll know where to look for new CRISPR systems. Knowing where to look should increase the rate of analysis of new systems, and with new discoveries like CPF1 happening, we can hope to find even more efficient CRISPR systems for practical application.

References:

1.

Rath, D., Amlinger, L. & Rath, A. The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie* **117,** 119–128 (2015).

2.

Barrangou, R. The roles of CRISPR–Cas systems in adaptive immunity and beyond. *In Innate immunity* (2015). doi:10.1016/j.coi.2014.12.008

3.

Universite Paris-Sud11. *CRISPRdb*. at <http://crispr.u-psud.fr/crispr/>

4.

* Zetsche, B., Gootenberg, J. S. & Abudayyeh, O. O. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* (2015). doi:http://dx.doi.org/10.1016/j.cell.2015.09.038