Javin deMello-Folsom
Biology 131
Spring 2018

Locating OriCs within a Genome

URL: https://repl.it/@jademello/Independent-Project

Motivation:

I was motived to take on this project, because I found the discussion of skew diagrams in class interesting. Moreover, when such an independent project was presented as a possibility, I immediately started trying to think of a way to solve it. Lastly, it has major implications for helping scientists address a real-life problem: finding the location of OriCs within a given genome after that genome has been sequenced.

Data Formatting:

Originally, the *S. solfataricus* genome was obtained from NCBI. However, this genome sequence proved too long to work with. So, three toy sequences were made by myself to be used as the working data. Toy sequences of varying complexity were made to text the functionality of the program. Toy1 was made to produce a skew diagram with fairly obvious OriCs, Toy2 was made to produce a skew diagram with slightly less obvious ones, and Toy3 was made to show a range of valleys used to test the input parameters.

Program Steps:

The program starts by obtaining the genome sequence from the .fasta file it's located in. Next, it creates a skew sequence of that genome sequence. With this skew sequence, it then calculates a backrange and skewdrop value (these are currently set as 15% of the sequence length and 20% of the total skew range, respectively). The program then uses a for loop to determine whether and given skew value within the skew sequence is a greater distance than the skewdrop from the highest value in the backrange. If so, this number (which is a potential OriC) is stored. These stored values are next separated out into groupings, based on if they're next to each other (within the same valley), and then the lowest number of each grouping is located. Next, the index of the lowest number in each grouping is found, and these numbers are stored and outputted as the potential OriCs. Finally, this program makes a skew plot and marks where the potential OriCs are.

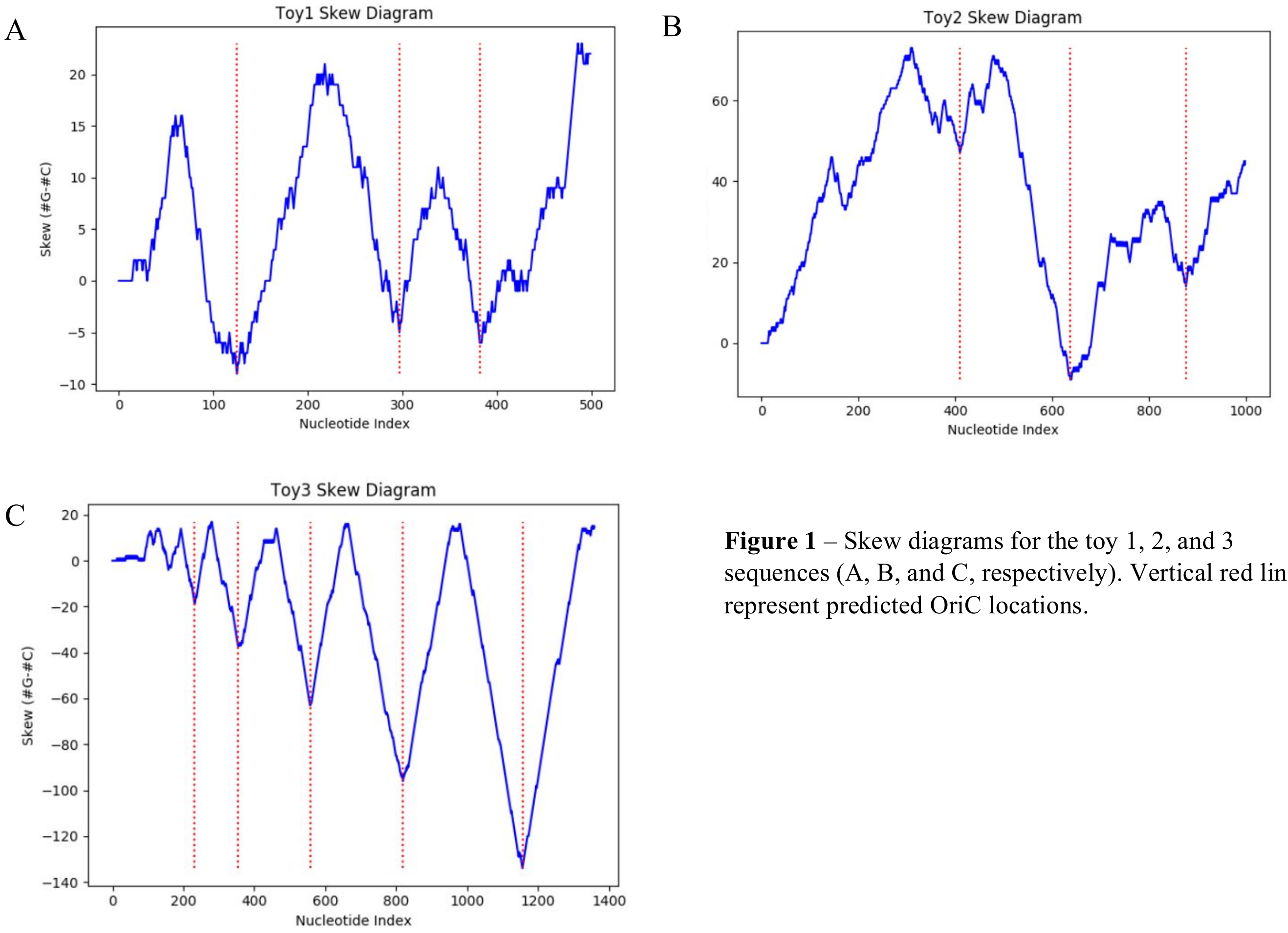

Figure 1 – Skew diagrams for the toy 1, 2, and 3 sequences (A, B, and C, respectively). Vertical red lines represent predicted OriC locations.

Results/Discussion:

The results showed that it is possible to predict possible OriCs, at least within the toy sequences. That being, given the somewhat arbitrary inputs for backrange and skewdrop, this program can successfully identify what would visually appear on the corresponding skew diagram to be a valley representing an OriC (Fig. 1A, Fig. 1B). However, given these predetermined parameters, there is some limit to size of the valley that can be detected. This is shown with the small valley just before the 200th nucleotide in Figure 1B and at the start of the sequence in Figure 1C.

Since the toy sequences 1 and 2 are vaguely mimic actual genomes, the assumption is that this program could potentially identify actual OriCs. Nevertheless, it would take a long time to complete a single run. Therefore, further tuning of the program would be necessary before applying it to a full-sized genome.