Ami Cooper

Identification of OriC Regions in Archaeal Genomes

1. Data were taken from the ncbi website. no manual trimming was done; just took the full genome file in fasta format as it was. Species is Sulolobus solfataricus.

2. **Part 1: Most Frequent Kmer in forward and rev complement**
   Input: string of genomic data
   Output: reverse complement of string and most frequent 9 and 13mers and the frequency of said kmers.

   Part 2: computing GC skew
   Input: string of genomic data (same as before)
   Output: list of values of GC skew ((C-G)/(C+G)) at each index and a plot of that list of values against the base pair number.

   Part 3*: sign changes**
   Input: list of values of GC skew
   Output: indices at which the GC skew changes from positive to negative
   **** this part was unsuccessful compared to the known value of where the paper says the origins of replication are.*

3. For the most part, my project was successful in that it identified and counted the 9 and 13mers in both the forward and reverse directions, and output a GC skew plot that looked kind of how it was supposed to. It was less successful in that the list of indices where the sign changes was a lot longer than i expected it to be, which makes it impossible to really determine from the output of the program where the origins of replications actually are (there are supposed to be 3 of them)

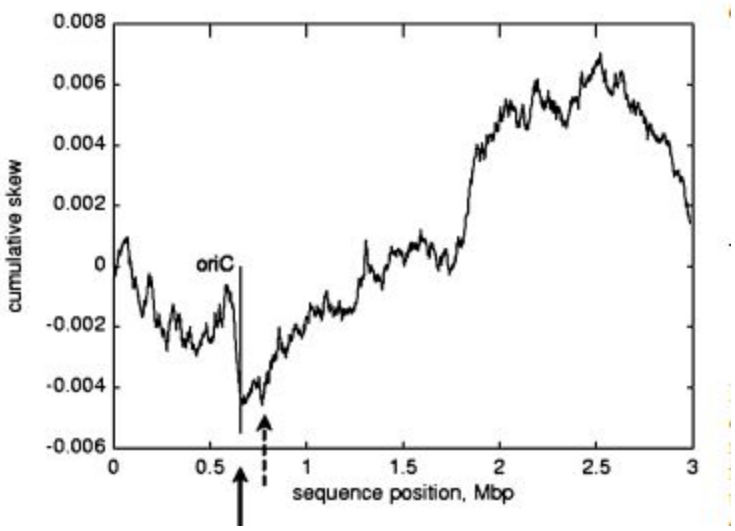

Figure 1.1 borrowed from Contrussi et al.'s Identification and autonomous replication capability of a chromosomal replication origin from the archaeon Sulfolobus solfataricus. Solid arrow indicates the origin of replication. However, another paper identifies two additional origins of replication that are not annotated by base pair location in that paper and therefore are not annotated here.

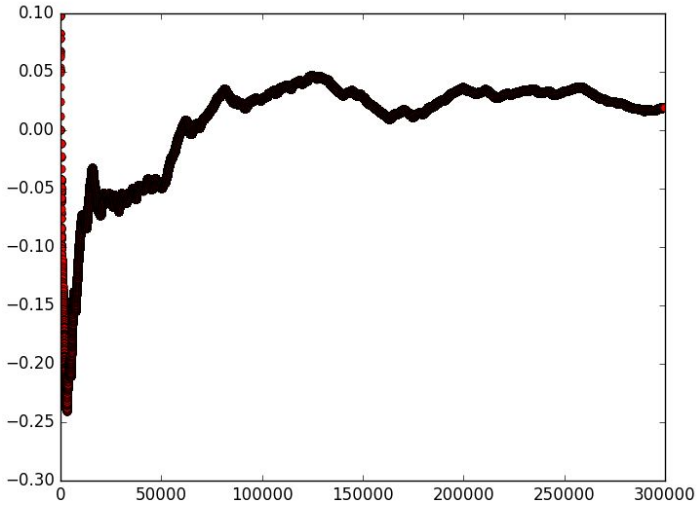

Figure 1.2 Plotted output of my python program Despite being on approximately the same scale, the output is completely different.

I conclude that my program is ineffective at predicting OriC regions in archaea and needs a lot of tweaking. If I had had more time, I would have also liked to see where in the

genome all of the kmers were located along the strand because their distribution might give a better sense of where the OriCs are.