Yurel Watson

Anna Ritz

Bio 131

3 May 2016

Identifying Origins of Replication Sites in Circular Genomes using a Scoring Method

In creating this program, I sought to develop a visual way of analyzing the possible origins of replication in circular genomes. While most of this legwork can already be done by simply looking at a skew graph, my program allows for the quick and efficient creation and storage of this skew graph data for later analysis. My program takes the highest and lowest skew values in a specified region in a genome and, from this, identifies possible origins of replication by scoring data points against the difference of these values.

The genomes included in this project were sourced from GenBank. Their identities and identification numbers are listed later in this document. In my program, the data is not altered in any way. The only modification made is when the genome is initially read; the program skips the first line as it is only a header and does not contain actual genetic information.

My program can be broken down into 4 sequential parts: Input, Input Processing, Printing and Analysis, and Exiting. In the Input section, the name of the .fasta file to be analyzed is entered and its genome is transferred to a string. Input Processing then iterates through this string and creates a list of **indexes** and a list of corresponding **skew** values. Note that the term *skew* denotes a running total of how many Gs or Cs have been read so far in a particular genome; Gs raise this score while Cs lower it. After this

processing is done, we enter the Printing and Analysis section. Here, the user is presented the final *skew* value and length of the *skew* list we just created – corresponding to the amount of nucleotides present in the genome. Now, a graph is created with the aforementioned

the origin of

Figure 1.1 Entire genome of Salmonella enterica serovar Typhi Ty2 (Accession: AE014613.1 GI: 29140506) with nucleotide

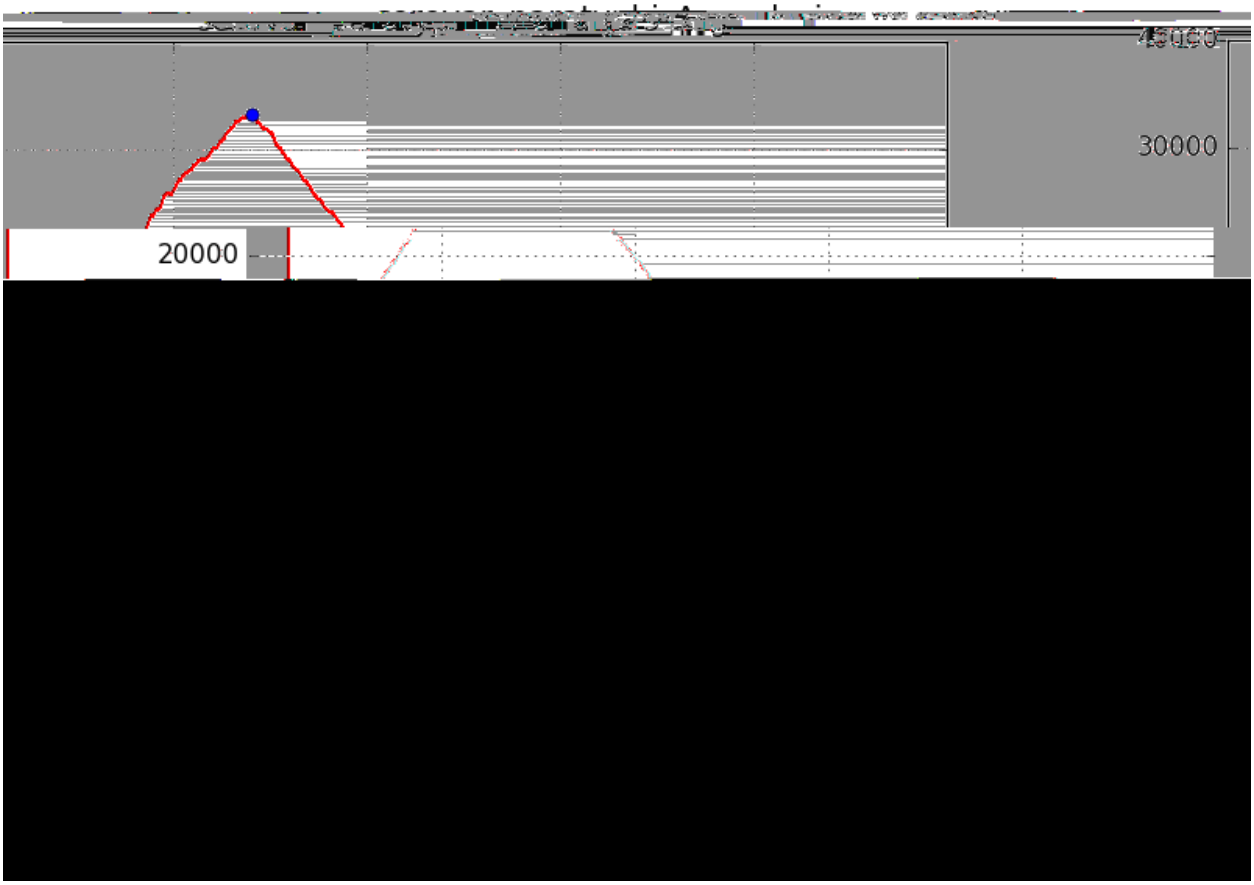

Figure 1.2 Entire genome of Salmonella enterica serovar Paratyphi A ATCC 9150 (Accession: CP000026.1 GI: 56126533) with nucleotide indexes and skew values plotted as x and y-values, respectively

Figure 1.3 Entire genome of