CS309: Computational Biology

Fall 2021

Lab 1: Searching for a Bacterial Replication Origin

Wednesday, Sept. 22, 2021

Name: Amna Khalid
Lab partner: Ebo Dennis

Introduction

DNA, or deoxyribonucleic acid, is the genetic material in humans and almost all other organisms. The data in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule and is called a nucleotide. Nucleotides are arranged in two long strands that produce a spiral called a double helix structure.

One of the unique qualities of DNA is that it is able to replicate itself. DNA replication is a semiconservative process i.e that each strand in the DNA double helix structure acts as a template for the construction of a new, complementary strand. In replication, one of the strands is oriented in the 3’ to 5’ direction (towards the replication fork), this is referred to as the leading strand. The other strand is oriented in the 5’ to 3’ direction, this is referred to as the lagging strand. As a result of their different orientations, the two strands are replicated differently. The leading strand is replicated continuously, while chunks of DNA, called Okazaki fragments, are then added to the lagging strand in the 5’ to 3’ direction.

Replication of the bacterial chromosome initiates at a single origin of replication that is referred to as *oriC* (Wolański et al., 2014). This process occurs via the combined operation of numerous proteins, including DnaA, which acts as an initiator. The binding of DnaA protein to its DNA binding sites—DnaA boxes—in the chromosomal *oriC* region is essential for the initiation of chromosome replication (Christensen, 1999). DnaA binds to short, typically 9 nucleotides long, segments within the replication origin known as a DnaA box. A DnaA box is a hidden message telling DnaA: “bind here!” DnaA can bind on both strands i.e both the DnaA box including its reverse-complement are equal targets. The reverse complement of a DNA sequence is formed by reversing the letters, interchanging A and T, and interchanging C and G.

In this lab, the origin of replication of bacterial organisms is examined to understand DNA and its replication process. Thus, for this lab, using different methods, the DNA replication origin or *oriC*, and the DnaA box of a bacterial genome will be located in order to answer the question: wherein a genome does replication begin?

Methods

The bacterial genome being used for these experiments is Deinococcus radiodurans strain R1 dM1 chromosome I, which is a radiation-resistant bacteria. The genome sequence is stored as a string using the readFile(fname) algorithm that reads in the sequence from a .FASTA file. To understand the genome, initially, the base composition percentages of the genome, as well as the localized base content of the genomes, are calculated. Further, the localized base content of the genome is plotted to determine the composition of the DNA sequence along with its bases and to determine if there are any interesting correlations that can lead to the identification of the DnaA box.

Further, to analyze whether the nucleotides guanine and cytosine are over-or under-abundant in a particular region of DNA, a GC skew is plotted. The algorithm thus traverses the genome, keeping a running total of the difference between the counts of G and C. Skew(seq) is defined as the difference between the total number of occurrences of G and the total number of occurrences of C in the first i nucleotides of the Genome. The skew diagram is defined by plotting Skew(sequence) as i ranges from 0 to |Genome|, where the initial skew value is set equal to zero. Using these values and solving for the Minimum Skew value on the graph, provides an approximate location of *oriC*.

In continuation of the search for DnaA boxes in the genome, frequent words in the *oriC* region are considered candidate DnaA boxes as these words might be the hidden message in the genome that starts replication. Thus, an algorithm freqkmers(seq, k) is designed to find the most frequent k-mers i.e. words with length k in the genome. The algorithm checks whether the k-mer appears in position 0 of the genome, position 1 of the genome, and so on. A set of frequent k-mers and their frequencies are stored and returned. However, this method does not always end up giving us an answer to where might the DnaA box be located since there are various frequent k-mers in a genome.

To further be able to identify the DnaA boxes in the genome, the most frequent k-mers with mismatches and reverse complements are identified. This algorithm mismatchesandrevcomplements(seq, k, d) takes in the number of mismatches d as a parameter by which it determines how many mismatches in a k-mer are present. The algorithm neighbors(dna, d) finds and stores the neighbors of the specific k-mer while having d mismatches. These mismatches are calculated using the hamming(a, b) function that calculates the Hamming distance between the two values, and the reverse complements of a sequence using a function called revcomplement(seq) are also determined. From the lists of k-mers and their reverse complements and the list of neighbors (with up to d mismatches), the k-mers with the max frequencies are returned. This method leads to the identification of a k-mer and its complement that can be the DnaA box for the genome. Specifically, the most frequent 9-mers (with 1 mismatch and reverse complements) are identified within a window of length 500 starting at the minimum position of the skew (which is the region suggested by the minimum skew as *oriC*).

To conclude, the DnaA boxes are narrowed down to one 9-mer and its complement via the use of different methods listed above. These methods are also used as a way of verification to confirm that the resulting 9-mer is indeed the DnaA box located near the *oriC* region of the genome.

Results

Initially, the genome sequence of the Deinococcus radiodurans strain bacteria is investigated via base composition analyses of the genome. Results of the base count and base percentages in the genome are shown in figure 1. Looking at the results, it is inferred that the G-C content of the bases is higher in comparison to the A-T content. Overall base C is shown to have the highest base composition in the genome followed by G.

Base	Base count	Percentage
A	435608	16.45%
C	888886	33.57%
T	436239	16.48%
G	886805	33.50 %
Total	2647538	100%

Figure 1

Further analysis of the genome composition is performed via looking at local base compositions of the genome using different window sizes (the length DNA sequence being analyzed), and intervals (the amount by which we move the window size). The graphs for window sizes of 20 Kb and 90 Kb are shown in Figure 2(a) and Figure 2(b) respectively.

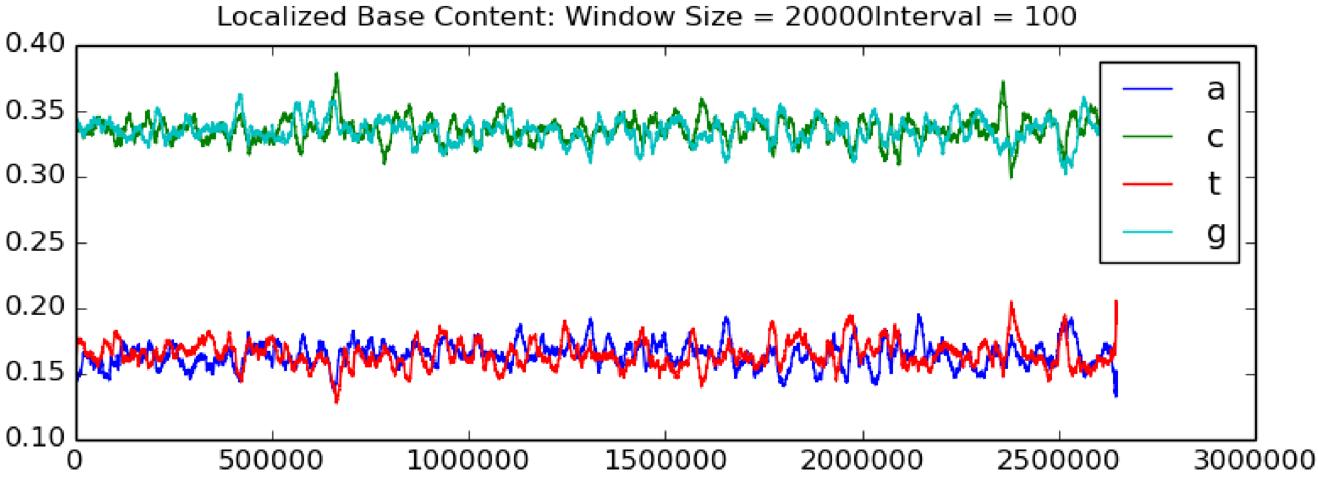

Figure 2(a)

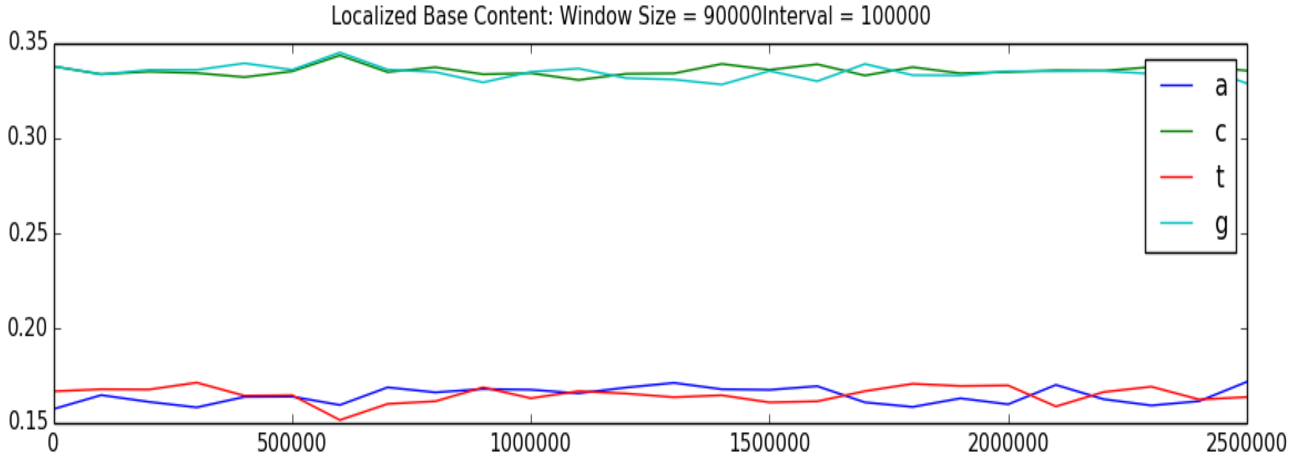

Figure 2(b)

Studying the separate graphs for each of the window sizes, it is evident that having a smaller window size gives more definitive values while having a larger window ‘smoothes out’ the patterns or genomic features. Similarly, the the 90k window interval is unable to show specific peaks of data whereas in the 20k window size, extreme values for different bases are evident and noticeable to observe fluctuations of the base compositions in the genome. In contrast, decreasing the window size more makes the graph overwhelmed with values making it hard to interpret.

While searching for *oriC*, one of the methods used is a GC skew. In equilibrium, both cytosine (C) and guanine (G) are present in equal frequencies. However, during the replication process, the leading strand contains more guanine than the lagging strand, hinting to the fact that the origin of replication is present nearby. Figure 3 displays the skew plot for the Deinococcus radiodurans strain of bacteria.

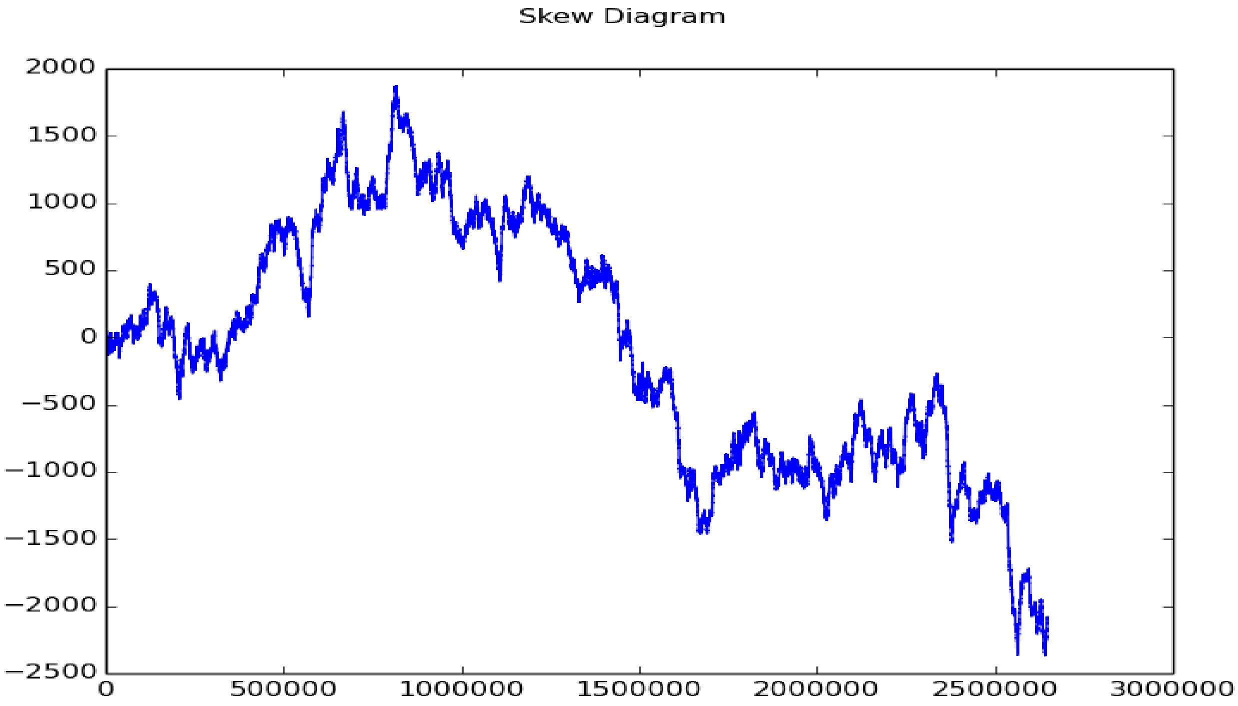

Figure 3

Further analysis of the skew gives us the maximum and minimum values where the max is at 815276 and the minimum is at 2562296. The 2562296 position in the genome is the approximate location of the *oriC* region in the bacterial chromosome. However, in an attempt to confirm this hypothesis, hidden messages or k-mers are searched for in the region as potential DnaA boxes.

Using the algorithm mentioned above, the most frequent 9-mers generated in the genome sequence is GCGGCGGCG. Specifically looking for frequent 9-mers near the *oriC* region with a window length of 1000bp, the function freqkmers(seq[2562296-500:2562296+500], 9)) returns 'CGACCTGAC', 'CGCTGACCG', 'GCGCTGACC', 'CCGACCTGA', 'CCGGCTCGG', and 'TCTGTCCAA' as the most frequent k-mers. Narrowing the window size to 500bp starting at position 2562296 of the genome, returns 'CGCTGACCG' and 'TCTGTCCAA'.

In an attempt to find DnaA boxes still, the most frequent 9-mers with mismatches and reverse complements in the region suggested by the minimum skew as *oriC* are found. The funtion returns teh most frequent 9-mer with 1 mismatch TGGTCGACC, along with its reverse complement GGTCGACCA.

Discussion

The goal of this lab was to identify the *oriC* region i.e. the origin of replication in a radiation resistant bacteria: Deinococcus radiodurans strain. Further, the lab also aimed to identify the DnaA box of the genome. Initially, the approach was to look at the base compositions and identify regions where there are discrepancies in the base compositions, However, the base composition profile of the genome does not reveal as much relative evidence to get a location of *oriC*. Observing figure 2(a), the changes in values of guanine and cytosine to be able to locate a higher content of guanine is not possible.

Similarly, continuing our search for *oriC* in the genome, a GC skew graph of the genome is projected. In this plot, a positive deviation from C corresponds to lagging strand and negative deviation from C corresponds to leading strand. Furthermore, the site where the deviation sign switches corresponds to the origin or terminal. A minimum value on the skew represents a higher content of guanine than its counterpart cytosine, leading us to conclude that the minimum value position is near the replication region. For this particular genome, the minimum rests at the 2562296 position giving an approximate localization of *oriC*.

To locate the DnaA boxes in the genome, the initial approach is to find frequent 9-mers in the genome sequence. However, this approach does not yield a conclusive result since it gives two 9-mers 'CGCTGACCG' and 'TCTGTCCAA', neither of which are complements of one another. This led to modifying the algorithm to finding frequent k-mers with d mismatches and their reverse complements.

The algorithm to find frequent 9-mers with upto 1 mismatch in the sequence returns CGACCTCAC and GTGAGGTCG. Notice that GTGAGGTCG is the reverse complement of CGACCTCAC, and since the presence of a k-mer and its reverse complement in a short region of a genome is statistically unlikely, now experimental results confirm that CGACCTCAC is the DnaA box in the Deinococcus radiodurans genome. However, each genome has a different DnaA box, nonetheless the same methodology coud be used to locate the *oriC* region as well as the DnaA boxes of other genomes.

External Sources

Christensen, B. B., Atlung, T., & Hansen, F. G. (1999). DnaA boxes are important elements in setting the initiation mass of Escherichia coli. Journal of bacteriology, 181(9), 2683–2688. https://doi.org/10.1128/JB.181.9.2683-2688.1999

Wolański, M., Donczew, R., Zawilak-Pawlik, A., & Zakrzewska-Czerwińska, J. (2015). oriC-encoded instructions for the initiation of bacterial chromosome replication. Frontiers in microbiology, 5, 735. https://doi.org/10.3389/fmicb.2014.00735