Searching for the Replication Origin for *Abyssicoccus albus*

Author: Kairuo Yan

Group Partner: Yanni Guo, Xinduo Fan

Date: September 16, 2021

Introduction:

In this lab, we explore a bacteria called *Abyssicoccus albus*. We examined the DNA of the bacteria and searching for its replication origin. *Abyssicoccus albus* have a DNA that is 1,745,789-bases long, which is in a relative smaller size compared to most Eukaryote. In the process of DNA replication, the proteins unwrap the DNA helix and build the complementary base pairs to copy original DNA. DnaA protein binds to DNA strand and contains a DnaA box that initiate DNA replication. Scientists have found the most frequent gene-bases pattern that appeared in the DNA strand is most likely to be the DnaA box. The change of DNA bases number also reveals the position of replication origin. We used computer science knowledge to facilitates the process of finding the origin. To find the most likely origin pattern or position, we would apply algorithms to move along the DNA strand and count the number of bases. Therefore, we would solve following major problems: how the change of the number of DNA bases could tell the replication origin, the most likely pattern for DnaA box, and the most likely DnaA box pattern with mismatches.

Methods:

We designed five major algorithms to find the pattern and position of the replication origin. In first algorithm, we get more knowledge about *Abyssicoccus albus* by counting the number of A, T, C, G bases in its DNA. The parameter is DNA sequence. In the second algorithm, which is called localizedBaseContent, and we have parameters as DNA and windowsize. we plotted the graph that show the percentages of four base composition in a sliding-window range throughout DNA strand. In the given window size, every time the window slides one base along the DNA strand, there's a base loss and a new base gained. And in the function, we will figure out what

those two bases are, and recalculate the base composition along the way that window slides down the DNA strand. Thirdly, we plotted a skew diagram that return the position of the replication origin with the minimum skew value. We calculated the difference between base G and base C. The parameter variable is DNA. According to the molecular biology, the turning point that the decreasing difference increases is the position of replication origin. The detailed biological concepts of the skew concept will be explained in the result part. The fourth algorithm aims at finding the most frequent k-mers in DNA strand. k-mers are base pattern with length of k. The function has four parameters, which are DNA, k, window_position, and windowsize. k is the number of bases in the DNA sequence; window_position is the position of the base window we are analyzing and windowsize represent the size of the window. The function first assigned counts to all frequent appeared pattern in k length in a dictionary. As the window moved one base along DNA, the algorithm repeated. Then we get the pattern with the most counts from the dictionary. The last algorithm aims to find the most frequent k-mers with mismatches. We first use mismatches (pattern, d) function to find the pattern with d number of mismatched bases. The pattern variable is the most frequent pattern. Then we get the reverse complement of the most frequent pattern. In the third function mostFreqWithComplement, we found the most frequent pattern in the DNA and plugged it back to the previous two functions for mismatches and reverse complement. We called these two functions and get the count of patterns in a dictionary. Eventually pattern with maximum count were found. It represents the most frequent pattern with one or two mismatched bases.

Results:

The size of original DNA strand is 1,745,789, and the composition of base A is 571392, the composition of base T is 579705, the composition of G is 293230, and the composition of C is 301462. We found in original DNA, base A and T are the most, and they are close in numbers. G and C have close amount and they are relative smaller in size compared to A and T. Then, we try to localize base content in the DNA. We designed a function that counting the percentage of each base in each 100 words sliding windows. Each time the window moved one base along the DNA, we will recalculate the composition of bases in that window. Eventually, we plotted line graphs showed how composition of ATCG changes along the way that given window sliding down the DNA strand. As results, for the base composition percentage in each 20,000 base windows, we found A and T take the higher percentages of base composition in each sliding window, while C and G take smaller percent of base composition in each sliding window. Initially A is more than T, and G is more than C. In around the middle position where the DNA replication direction changes, T gets more than A, and base C gets more than G (figure 1). For the base composition percentage in each 90,000 base windows, the results are same as the 20,000 base windows, and we can be easier to see the comparison and changes in amount for A, T, C, G composition percentages (figure 2).

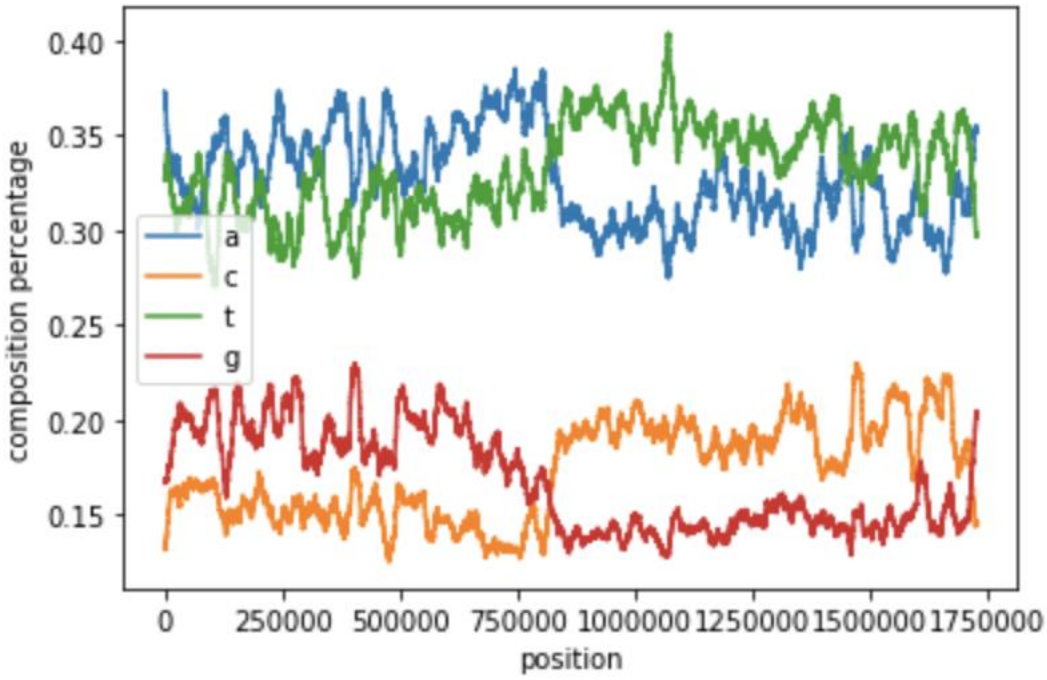

Figure 1. The Base Composition Percentage in each 20,000 base sliding windows. The composition percentage of base A and T is overall higher than base C and G. Initially along the DNA, A is more than T, and G is more than C. In around the middle of DNA, base T gets more than A, and base C gets more than G.

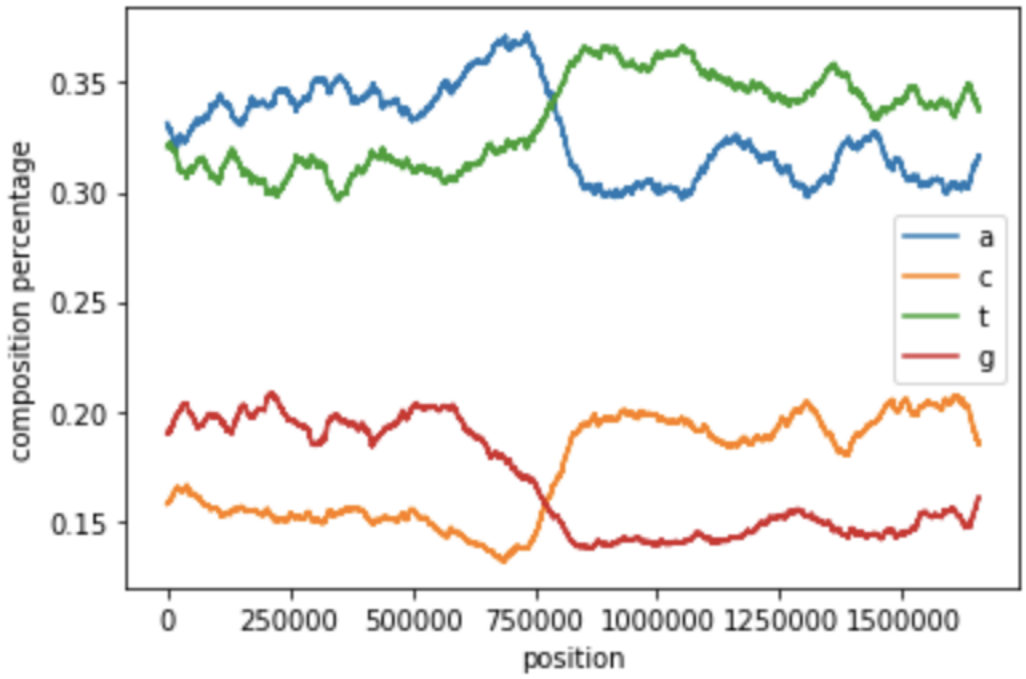

Figure 2. The Base Composition Percentage in each 90,000 base sliding windows. The composition percentage of base A and T is overall higher than base C and G. Initially along the DNA, A is more than T, and G is more than C. In around the middle of DNA, base T gets more than A, and base C gets more than G.

In the third experiment, we designed an algorithm that displays the skew diagram of DNA and returns the position with minimum skew value. We looked for the turning point where the decreasing difference between number of G and C increases (Figure 3). Then we zoomed in the graph and find the turning point is between position 1,720,000 and 1,730,000 (Figure 4). During the DNA synthesis, the direction is always in the 5' to 3' direction. While in replication, one strand copied in a forward way, and another strand copied in a reverse way. The strand copied in the reverse way tends to mutate base C and decreases its amount (Boyle, 1991). To find the position where replication starts, we could build algorithms to calculate the difference between

the number of base C and G in the replication. If there's a turning point where the difference changes from decrease to increases, the turning point would be the position of replication origin. For *Abyssicoccus albus*, the origin position is 1723790.

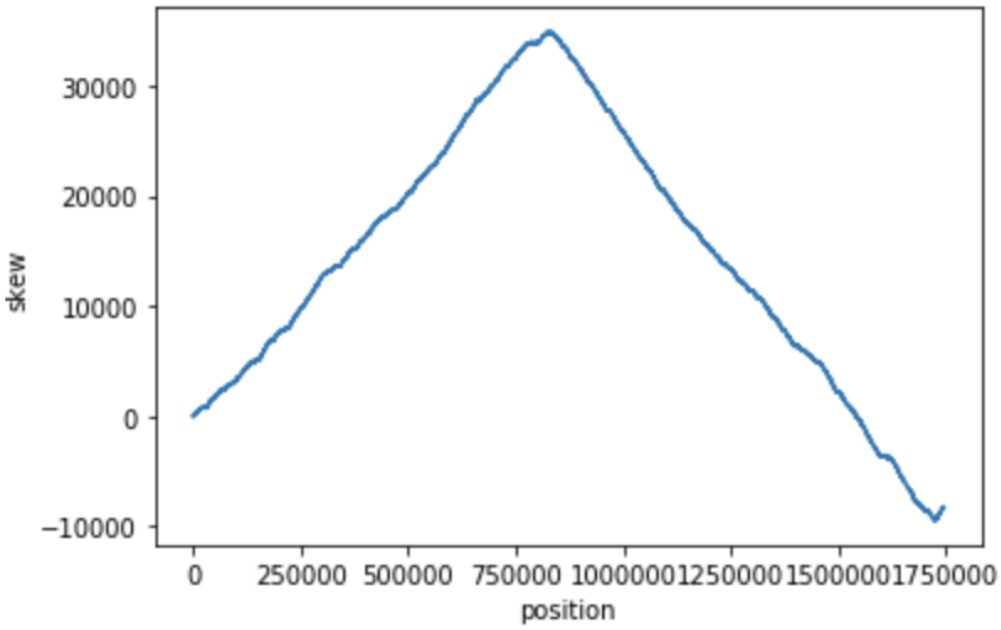

Figure 3. Skew Diagram. The skew initially increases and reach the peak in around the middle position of DNA. Then it decreases till it's lowest points, which is in the range between 1,600,000 and 1,750,000.

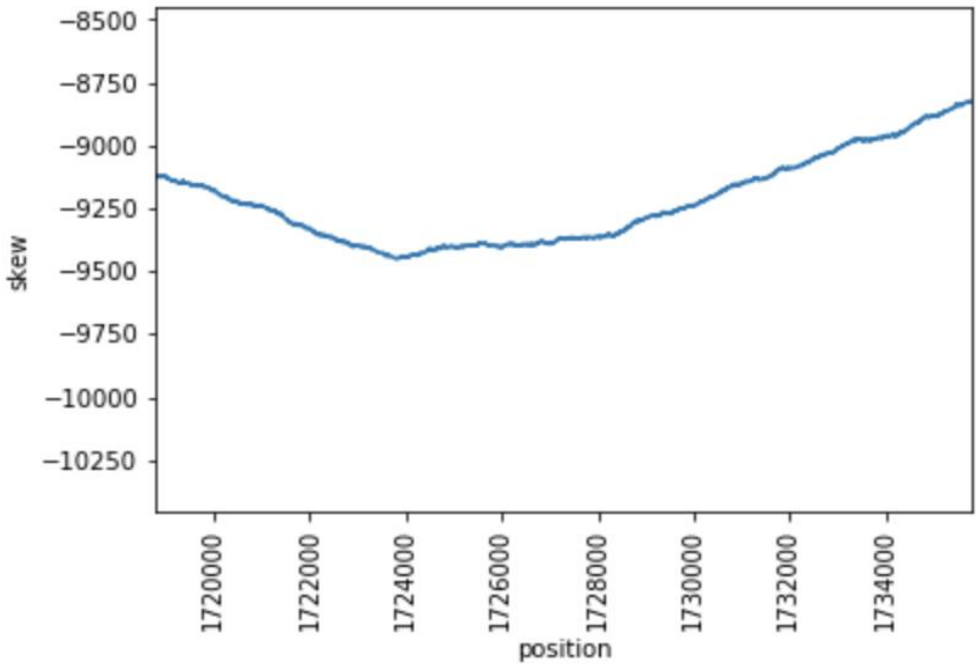

Figure 4. Zoomed-in Skew Diagram. The zoomed-in skew diagram for the lowest point and it is in the range between 1,720,000 and 1,725,000.

In the fourth experiment, we designed an algorithm to find the most frequent pattern that appears in the DNA strands. We use algorithms to get all patterns and the count of these patterns. Eventually, we get the pattern with the maximum count for the most frequent pattern. We are looking for the most frequent k-mers is because DnaA protein binds to DNA strand to a DnaA box that initiate DNA replication. Scientists have found that the most frequent showed gene-bases pattern that appeared in the DNA strand is most likely to be the DnaA box (Fu, 2021). As results, the DnaA box for the entire sequence are mostly to have a 9-mer pattern as 'TTTTATTTT'. The DnaA box centered at the position of the minimum skew where we analyze in each 500bp window is mostly to have a 9-mer pattern as 'TGTGGATAA'.

In the last experiment, we designed three algorithms to find the most frequent pattern with mismatches and its reverse complements that appears in the DNA strands. As a result, we found the most frequent 9-mers with one to two mismatches in the 500 bp windows are 'ATGTTTTTT', 'TTTTTTGAA', 'TTTTTTTGA', 'TATGTTTTT'. The reason for finding the mismatches and reserve complements is because during the replication, the mutation also happens. So, when we set the number of mismatches, we keep counting the possible mutated patterns so that we could locate the replication origin more accurately. And it is the same purpose that we count the reverse complements of the most frequent patterns.

Discussion：

In our experiment, we provide a visual insight on the composition percentages for base A, T, C, G. We found the amount of base A and T is always larger than C and G. Based on further data analysis, our result provides us the information of the replication origin position and the possible patterns of DnaA box. The algorithms allow us to change the variables so we could get the result for different conditions. We found that the replication origin for the entire DNA strand is 1,723,790, and the most frequent 9-mer pattern for entire sequence is ‘TGTGGATAA’. Since DnaA box locates at the minimum skew, we also found that the most frequent 9-mers in the 500 bp window centered at the position of minimum skew, and the result is pattern 'TTTTATTTT'. During the replication process, the DNA polymerase make mistakes along DNA strand when building the corresponding base pairs (Pray, 2008). Therefore, our algorithm sets the tolerance variables for base mismatches. We found the most frequent 9-mers with one to two mismatches in the 500 bp windows are 'ATGTTTTTT', 'TTTTTTGAA', 'TTTTTTTGA', 'TATGTTTTT'. In this research, the major difficulties are understanding the principles of DNA replication and

implement them into algorithms. We also took a huge amount time trying to figure out the logics of algorithms, especially on the algorithms that running through entire DNA sequence within sliding windows, and the one for finding mismatched pattern. There are some questions arose in this and further research. For example, when we are looking for the most frequent patterns in DNA strands, we assume the 9-mer pattern would be the most appeared pattern. However, there are possibilities that if we reset the k to other length, and there exists other patterns where DnaA box locate. There is a 13-mer DnaA boxes for E. colioriC region, which contains the consensus sequence of the DnaA boxes and two flanking bases (Fujikawa, 2003). Thus, I think it would be helpful if we could use methods to determine the most accurate length of frequent pattern. Also, in our research, we are looking for a prokaryotic genome, which is much less in the size compared to eukaryotic genome. Additionally, researchers found that there's a single replication origin for most Prokaryotes since they usually contain a small circular chromosome (Barry, 2006). However, eukaryotic genomes could have multiple replication origins due to their size and linear shape (Méchali 2010). Therefore, for further research, there could be possibility that the length of sequence inhibits our process of data analysis, and that is a challenge I hope to learn how to solve.

Bibliography:

Barry ER, Bell SD 2006. DNA replication in the archaea. *Microbiol Mol Biol Rev* 70: 876–887

Boyle WJ, Smeal T, Defize LH, Angel P, Woodgett JR, Karin M, Hunter T. Activation of protein kinase C decreases phosphorylation of c-Jun at sites that negatively regulate its DNA-binding activity. Cell. 1991 Feb 8;64(3):573-84. doi: 10.1016/0092-8674(91)90241-p. PMID: 1846781.

Fu, H., Redon, C.E., Thakur, B.L. *et al.* Dynamics of replication origin over-activation. *Nat Commun* **12,** 3448 (2021). https://doi.org/10.1038/s41467-021-23835-0

Fujikawa N., Kurumizaka H., Nureki O., Terada T. Structural basis of replication origin recognition by the DnaA protein. *Nucleic Acids Res.* 2003; 31:2077–2086. doi: 10.1093/nar/gkg309.

Méchali M 2010. Eukaryotic DNA replication origins: Many choices for appropriate answers. *Nature Rev Mol Cell Biol* 11: 728–738

Pray, L. (2008) DNA Replication and Causes of Mutation. *Nature Education* 1(1):214