CS 309: Computational Biology

Fall, 2021

Lab 1: Searching for a bacterial replication origin

Due Friday, September 17

For this lab, you will work in groups of 2–3. All members of each group are expected to contribute equally to all aspects of the lab. Upon completion of the project, each *individual* will write a report to summarize and analyze their findings.

Introduction

For this lab, you will attempt to locate the origin of replication (*ori*) in a chosen bacterial genome.

Experimental Objectives

1. **Choose a bacterial genome to study.**

   Do some research on different kinds of bacteria and choose one to study. If the genome has been sequenced, it can be found in the NCBI database at `https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/`. Download the bacterial genome in FASTA format. Make sure that you getting the sequence from a complete chromosome.

2. **Read the genome**

   Write a function that reads a FASTA file containing a single sequence and returns the sequence as a string. Your function should take the name of the FASTA file as a parameter. Use this function to read your bacterial genome for use in the steps that follow.

3. **Base content**

   Write a function that, given a DNA sequence, returns its size and its base composition in a dictionary. You can then use this information to find the frequency of each base.

   Note that additional symbols are used in genome data files when there is uncertainty in the data at a particular position:

Symbol	Meaning
`Y`	pyrimidine (`C` or `T`)
`R`	purine (`A` or `G`)
`K`	`G` or `T`
`M`	`A` or `C`
`S`	`C` or `G`
`W`	`A` or `T`
`B`	not `A` (`C` or `G` or `T`)
`D`	not `C` (`A` or `G` or `T`)
`H`	not `G` (`A` or `C` or `T`)
`V`	not `T` (`A` or `C` or `G`)
`N`	unknown base

4. **Localized base content**

   Write a function to determine and plot the base composition using a "sliding window" approach. Your plots should display the distribution of the base composition with the genome position on the x axis and base composition (as a percent) on the y axis. Run these analyses with window sizes of 20 Kb and 90 Kb.

5. **Skew diagram**

   Write a function that displays the skew diagram for a DNA sequence and returns the position with the minimum skew value (problem 1F on page 27). Be careful here — remember that these are circular chromosomes. So if the skew is decreasing up to the end of the sequence, look back to the beginning of the sequence to see if the decrease continues. Run your function on your bacterial genome. What does this tell you about where *ori* might be located?

6. **Most frequent k-mers**

   Write a function that returns the most frequent k-mers in a DNA sequence (problem 1B on page 8). Your function should take a DNA sequence and k as parameters. Use an efficient algorithm that requires only a single pass across the sequence.

   What are the most frequent 9-mers in the entire sequence? What are the most frequent 9-mers in the 500 bp window centered at the position of the minimum skew? Because there can be random variation in skew values, also look at several other 500 bp windows in that area.

7. **Most frequent k-mers with mismatches**

   As you read, it turns out that the DnaA protein can bind to approximate DnaA boxes as well as "perfect" ones. So we really need to be looking for groups of frequent k-mers that may differ in a few bases. And we also need to include their reverse complements because DnaA boxes can appear on either strand.

   So write a function that returns the most frequent k-mers with up to d mismatches in a DNA sequence and its reverse complement. In other words, return the k-mers that maximize the sum $\textsc{Count}_d(dna, pattern) + \textsc{Count}_d(dna, \overline{pattern})$ where *pattern* is the k-mer and $\overline{pattern}$ is its reverse complement. (This is problem 1J on page 31. Also make sure you read pages 48–51.)

   Finally, apply your function to find the most frequent 9-mers with 1–2 mismatches in the 500 bp windows that you considered previously.

Lab Report Requirements

Although the design and execution of the lab will have been done as a group, you will write your report, including the construction of figures and/or tables to present your results, *independently.*

Please refer to the separate *Lab report guidelines* document for instructions on formatting your report.

Please submit your report, the file(s) containing your program(s), and your genome FASTA file.