UTAH STATE UNIVERSITY, COMPUTER SCIENCE DEPARTMENT

CS 6670 Advanced Bioinformatics Assignment 3

Gopal Menon

July 18, 2016

1. **Biological Problem Description and Motivation**

   Regulatory motifs are sequences of nucleotides that lie upstream of genes in a DNA molecule [1]. Proteins known as transcription factors bind to these motifs and by doing so, encourage RNA polymerase to transcribe the downstream gene. Motif finding is the process of discovering such motifs without prior knowledge of what the motif looks like. The approach to motif finding is based on the assumption that frequent or rare words may correspond to regulatory motifs in DNA. If a word occurs considerably more frequently than expected, then it is more likely to be some sort of "signal".

   Since we may not know what a motif looks like, we cannot just do a search for a nucleotide sequence within a DNA sequence. We need to be able to find out frequently occurring nucleotide sequences that may indicate a regulatory motif. A further complication is that some regulatory motifs may have mutations at some nucleotide locations and we still need to be able to find them.

2. **Problem Definition**

   A formal problem definition is given below [1]:

   Median String Problem:

   Given a set of DNA sequences, find a median string.

   Input: A $t \times n$ matrix DNA, and L, the length of the pattern to find.

   Output: A string v of L nucleotides that minimizes $TotalDistance(v, DNA)$ over all strings of that length.

   In other words, given a set of t DNA sequences of length n and a target median string length of L, we need to find a median string v of length L nucleotides that minimizes

$TotalDistance(v, DNA)$. Here $TotalDistance(v, DNA)$ is the smallest hamming distance between the median string and all possible starting points starting points in each of the DNA sequences.

```
caggggcaggaagacagagcagctgacacttccagaaatagctggccaga
                gtagtaa
```

In the above string of nucleotides (the longer one could be thought of as a DNA sequence and the shorter one as a potential median string), the hamming distance at the position shown will be the number of nucleotides that are different. We can slide the potential median string and find the smallest hamming distance between it and the DNA sequence. The total distance will be the sum of minimum hamming distances between the potential median string and all the DNA sequences. The string of length L that minimizes the total distance will be median string.

3. **Results with Synthetic Data**

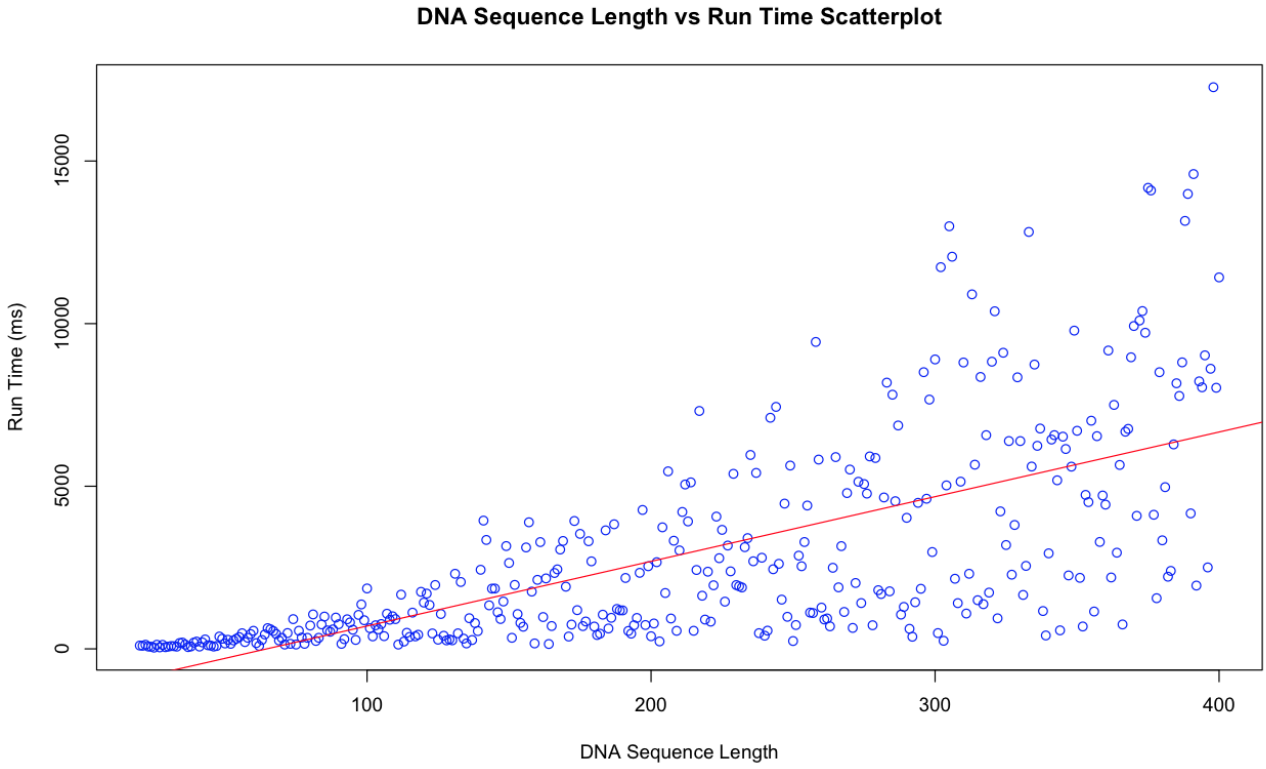

Figure 1: DNA Sequence Length vs Run Time Scatterplot

4. **Results with Real Data**

REFERENCES

[1] Jones, Neil C., and Pavel Pevzner. "Regulatory Motifs in DNA Sequences." *An Introduction to Bioinformatics Algorithms.* Cambridge, MA: MIT, 2004. N. pag. Print.

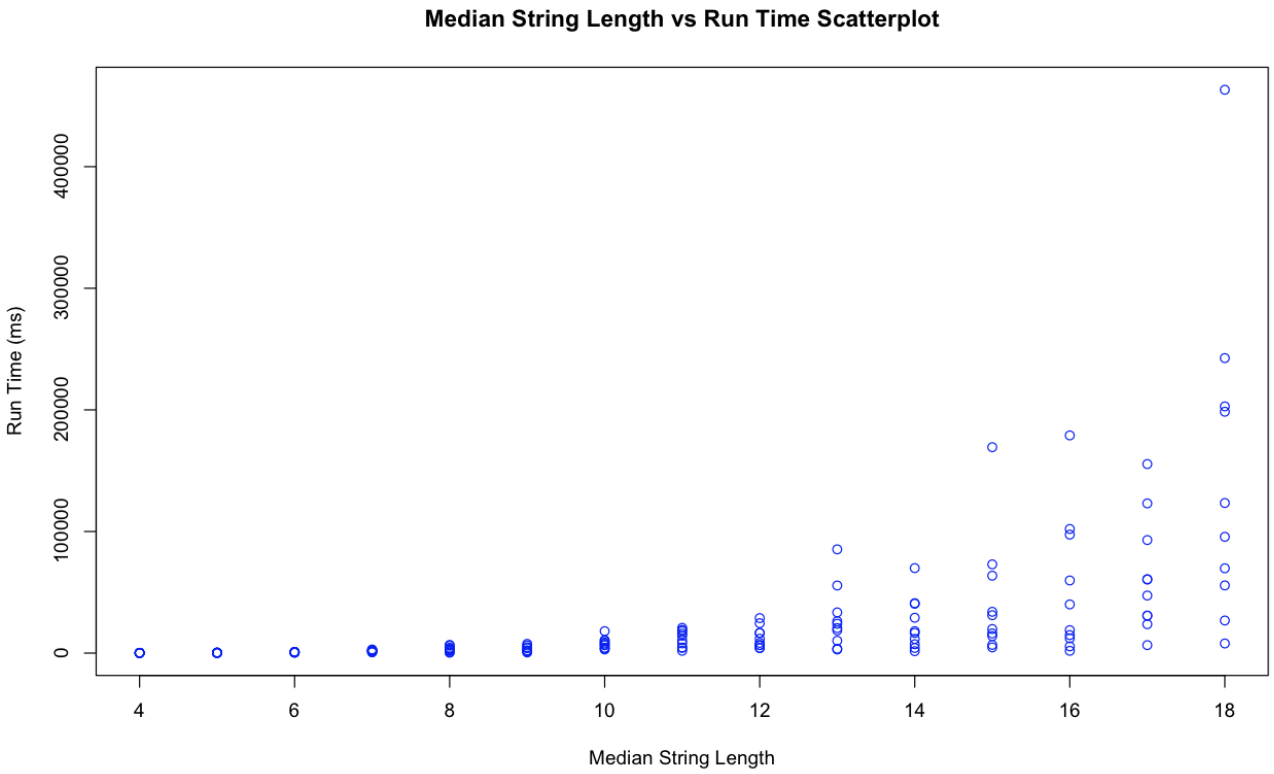

Figure 2: Median String Length vs Run Time Scatterplot

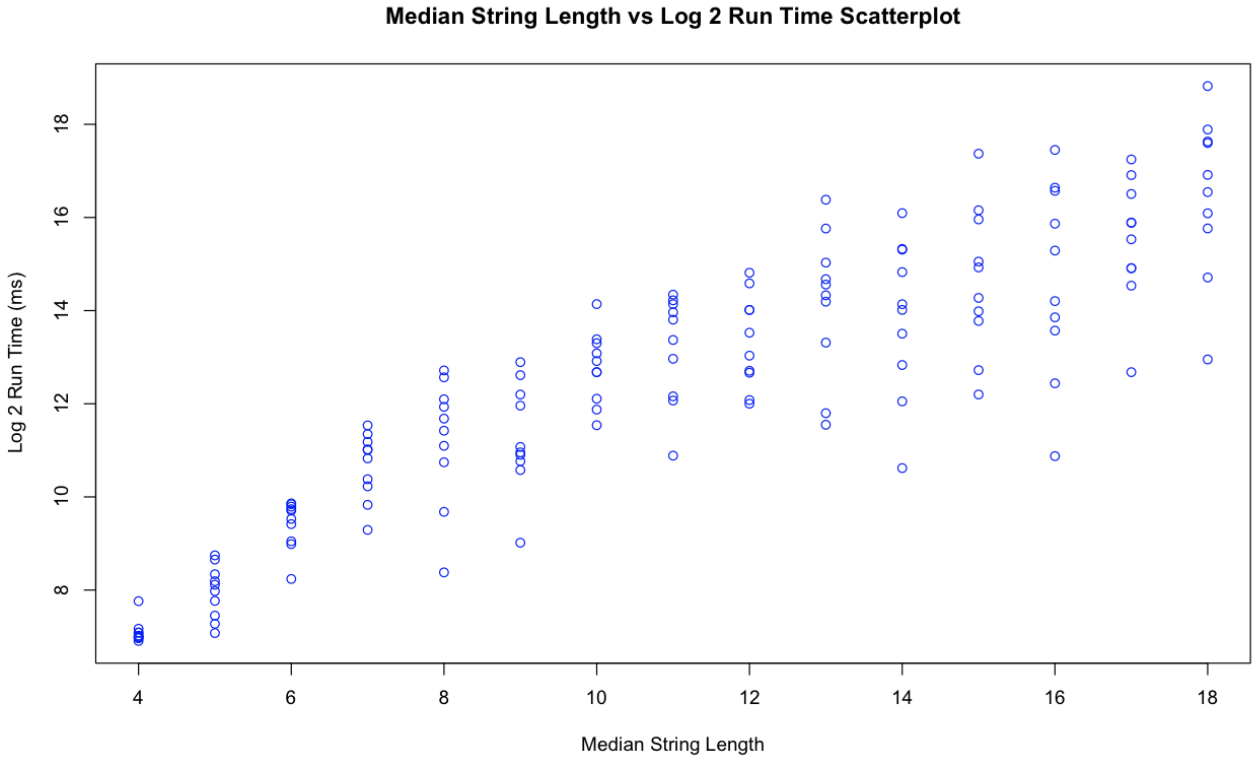

Figure 3: Median String Length vs log base 2 Run Time Scatterplot