CS 6670 Advanced Bioinformatics Assignment 3

Gopal Menon

July 20, 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

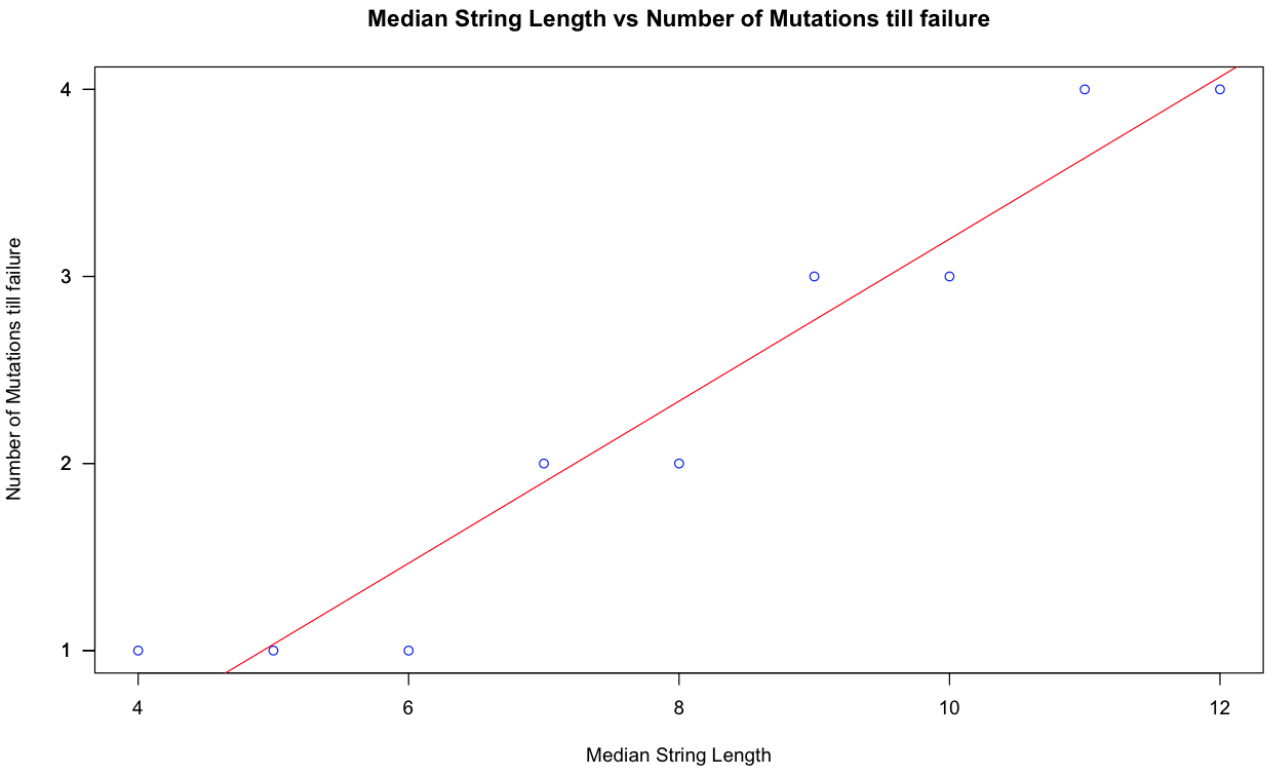

Figure 1: Median String Length vs Number of Mutations till failure

$TotalDistance(v, DNA)$. 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.

A median string described above would indicate a regulatory motif and its mutations that occur more frequently than expected, in a DNA sequence.

3. **Results with Synthetic Data**

   a) **Mutations till failure**

   Figure 1 shows the trend for the median string length versus the number of mutations at which the system fails to find the inserted median string. The test was done with 25 DNA sequences each of length 200. The blue circles are the data points and the red line is the line of best fit. It shows that with increasing median string length it takes more mutations for the system to fail to find the same median string that was inserted.

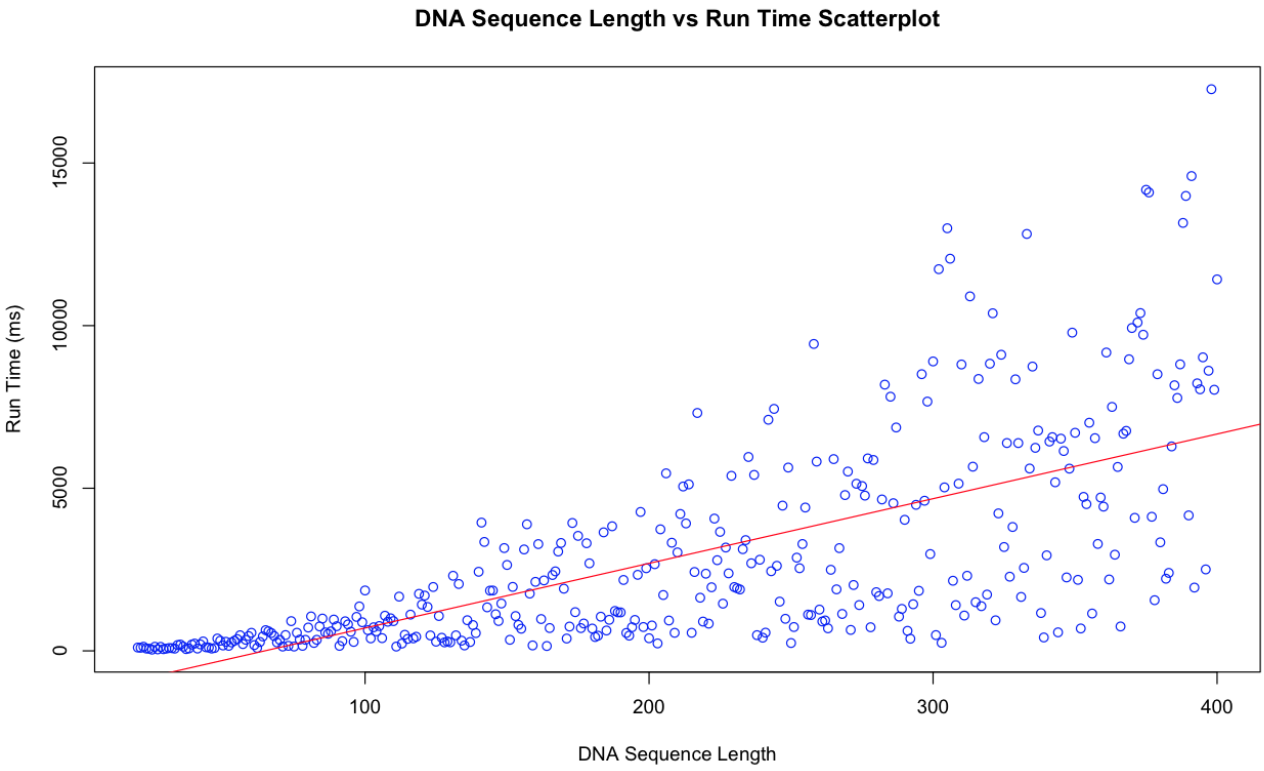

Figure 2: DNA Sequence Length vs Run Time Scatterplot

b) **DNA Sequence Length vs Run Time**

Figure 2 shows the plot for DNA Sequence Length vs Run Time for finding a median string. The data was collected for DNA Sequence Length varying from 20 to 400 nucleotides, 25 DNA sequences and a median string length of 8 with no mutations. For each DNA Sequence Length we can see data points varying from low to high run times. The reason is that the median string is randomly generated and those median strings that are checked for first in the branch and bound, are found earlier. The red line is the line of best fit and it shows a linear increase in run time with DNA Sequence Length.

c) **Median String Length vs Run Time**

Figures 3 and 4 show the plots for median string length vs run time to find a median string. The only difference between the two is that Figure 4 has a logarithmic scale for the run time. The plot seems to indicate that the run time is exponential in the length of the median string. The reason is that the logarithmic plot shows a linear relation.

4. **Results with Real Data**

Here are the results for the target length and the corresponding median strings found when run on the human DNA.

Target Length	Median String found
10	AAAATTCAAA
11	CTCCTGACCTC
12	TTTTGTATTTTT

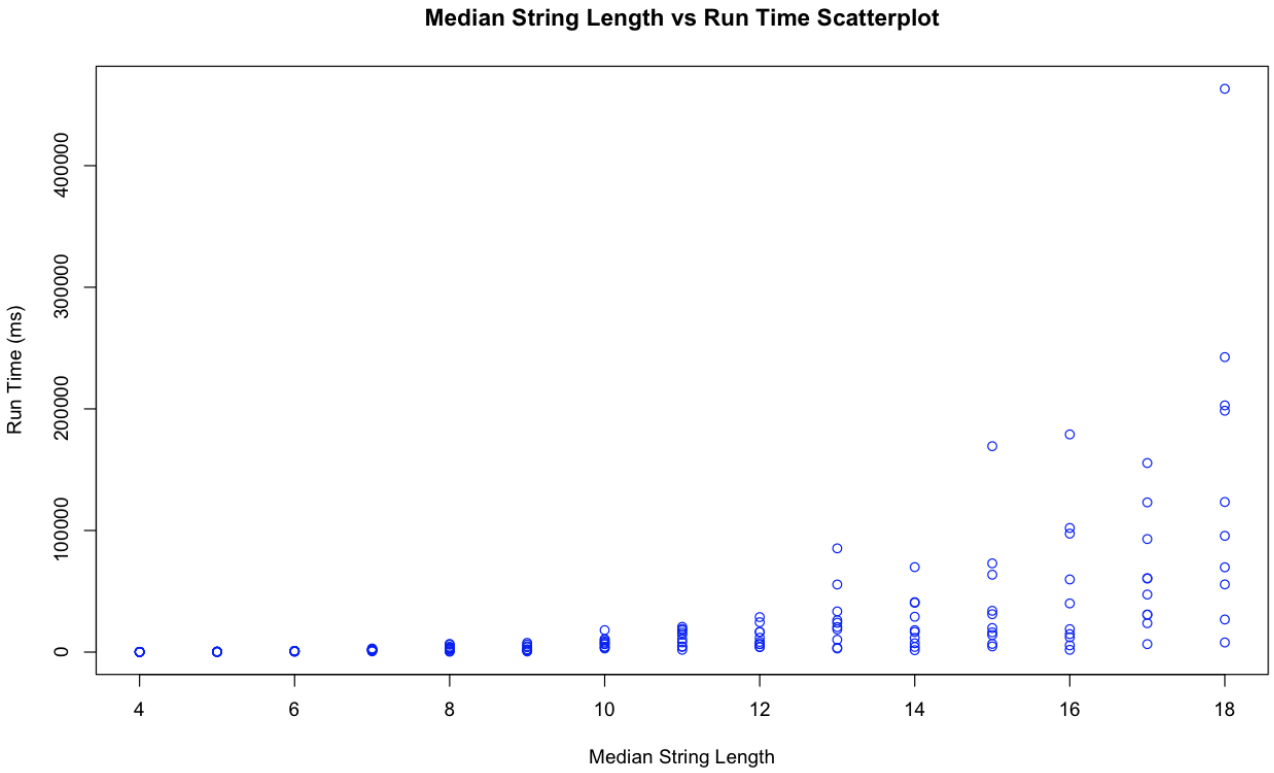

Figure 3: Median String Length vs Run Time Scatterplot

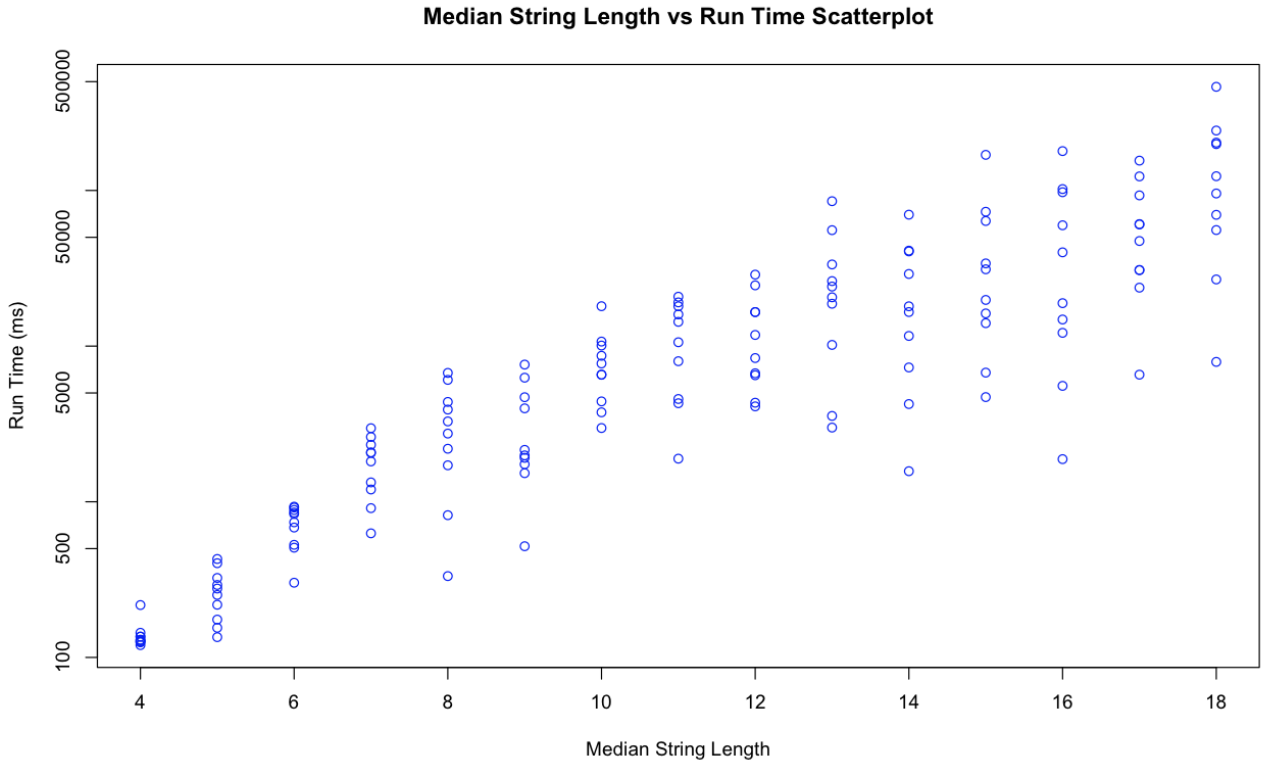

Figure 4: Median String Length vs Run Time Scatterplot

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.

A.
PROGRAM PARAMETERS

These are the run time parameters to be provided while running the program.

	Parameter Number				
Usage	1	2	3	4	5
DNA sequences in .txt files	T				
Use Human DNA sequences	H				
Use generated DNA sequences	G	DNA Seq Len	# of DNA Seq	Median Str Len	# of mutations
Repeat till median string not found	F				
Runtime with DNA sequence length	L				
Runtime with median string length	M				