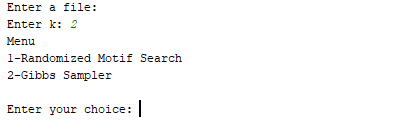
Computational Genomics Report

Homework #2

In this assignment, we search for motifs, and we try to find the consensus string. We implemented two algorithms “Randomized Motif Search” and “Gibbs Sampler”, we ran both algorithms and compared the scores and consensus strings obtained for different k values.

First, we have prepared an input file. Our input file has 10 lines and each line contains strings of length 500. Each line represents a randomly DNA string with 500 nucleotides. We have inserted our k-mer with 4 mutations in random positions into the DNA string.

Our program takes a file as input and a k value which will be the length of the consensus string.

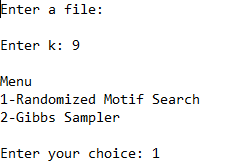


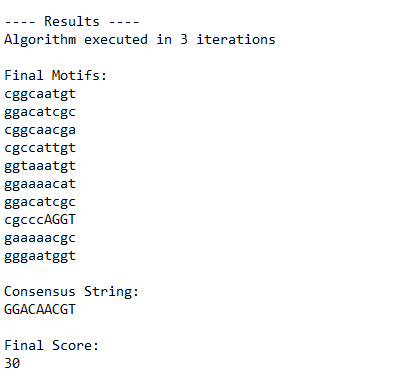
If we didn’t write an input file, the program will take the input file in the folder as default. If you press enter, the menu will pop-up. You can choose between two algorithms which are Randomized Motif Search and Gibbs Sampler.

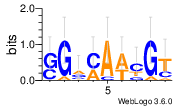
We will only put the screenshots of final iterations for each run. For each algorithm and different k values we will save a txt file in the project directory. Detailed information for each algorithm run will be visible in the output files.

Randomized Motif Search :

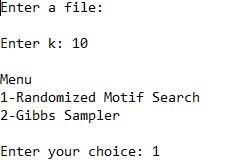
K = 9

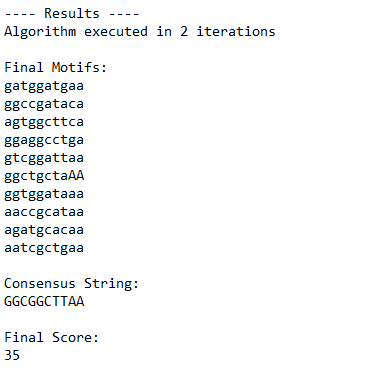


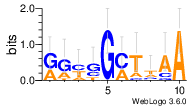




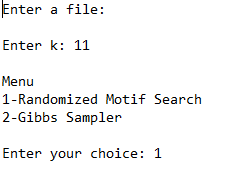
K = 10

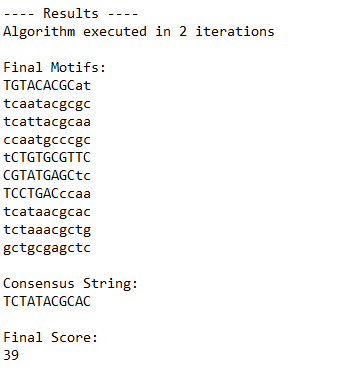


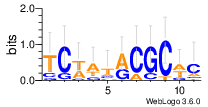




K = 11

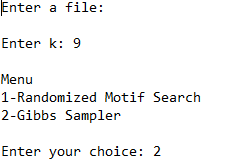


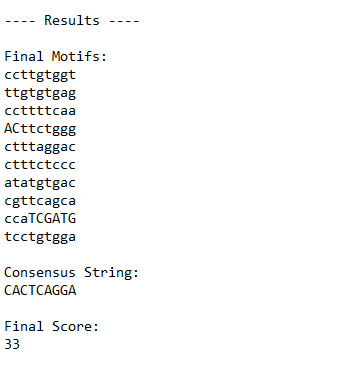


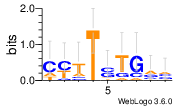


Gibbs Sampler :

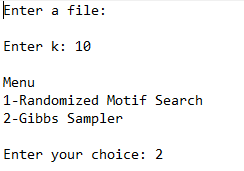
K = 9

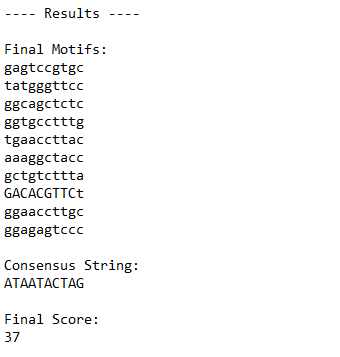


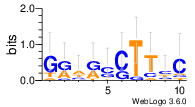




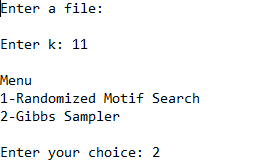
K = 10

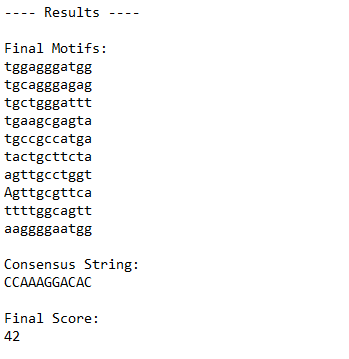


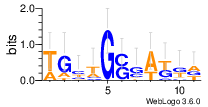




K = 11







Implementation Detail :

Randomized Motif Search:

In randomized motif search we start by selecting random motifs from each of the 10 lines and save them in an array called randomMotif. Randomized motif search algorithm implemented as “*randomized\_motif\_search(randomMotifs,k)*” takes two inputs such as randomMotifs and a k value. The k value indicates the length of each motif. The algorithm at initial runs with these random motifs and stops if the newly calculated score is greater than the current one. (The currently score is better then the new one). Randomized motif search algorithm firstly counts each nucleotide type for given motifs in column manner and from these values it constructs the consensus string. Later we calculate score, profile matrix and the from the profile matrix it iterates over each DNA string and gets the string with highest probability and selects this one as the new candidate motif. Doing this for each DNA string it creates the new set of motifs.

Gibs Sampler:

Gibs Sampler algorithm also starts by selecting random motifs from the DNA strings given in the input file. Firstly it calculates the score for these random motifs. Then the algorithm randomly chooses one row and doesnt takes this motif into count when doing the calculation for profile matrix. Unlike randomized motif search Gibs Sampler adds up one to each count of nucleotide in the count matrix. By doing so even if the nucleotide doesnt appears in the column it gives a very little probability in the profile matrix. After that the algorithm calculates each probability of the motifs by iterating on the DNA string which the algorithm has choosen to be discarded in the start of the iteration. Using these probabilities it chooses one motif randomly but ofcourse the higher probability has higher probability to be choosen. Gibs Sampler algorithm runs until the newly calculated scores are not better then the current score for 50 iterations. Gibs Sampler algorithm moves in small steps to obtain the consensus string.