In this project, we implemented two motifs finding algorithms Randomized Motif Search and Gibbs Sampler. And with the help of these algorithms, we tried to find consensus string from given DNA sequences.

Input File Creation and Insertion of K-Mer Motifs

First, we created input file which contains the 10 DNA sequence each with the length of 500 characters long generated by using https://www.bioinformatics.org/sms2/random\_dna.html website. We inserted 10-mer motif AAAAAAAAAA with 4 mutations in random positions into

10 DNA sequences.

10-mer motifs with 4 mutations as follows:

ACCAATAGAA

AACTGAACAA

GACAATACAA

AAGACTAATA

ATAGACAGAA

GAAACATAAT

ACAAGAGAGA

AATTAGAAAC

AATGGATAAA

AGAATGCAAA

Randomized Motif Search

Code snippet below is from the Main method where we start to execute our algorithms. We run the randomized motif search algorithm 5000 times. Because randomized motif search ends when it does not enhance the score. It means that it might stuck at local optimum. Thus,

We run the algorithm again and again to obtain the best consensus k-mer.

And then we put the returning best motifs into the HashMap with its score. After that we sort the HashMap ascending order by score value. And we select the lowest score value motifs which are the best motifs among all experiments.

Text

Description automatically generated

The method inside the below shows the implementation of the randomized motif search. First, we select the random k-mer motifs from sequences. And then, we create profile matrix from these motifs. We select new k-mer motifs with high score depending on calculation by using profile matrix. Finally, algorithm ends when it cannot improve the motif score and return the best motifs.

Text

Description automatically generated