CSE 4065 – Computational Genomics

Programming Assignment # 1

In this assignment, you are going to search for motifs, and try to find the consensus string.

You will implement Median String algorithm, Randomized Motif Search and Gibbs Sampler, run all algorithms and compare the scores and consensus strings obtained for different k values.

Input File:

- Prepare an input file which has 10 lines, where each line contains strings of lengt 500.
- Each line will include a randomly generated DNA string with 500 bases.
- Insert a 10-mer with 4 mutations in random positions into the DNA string.
- You will have total of 10 DNA strings, so you will use 10 motifs and find the consensus string.

Details:

- Select a random position between 0 – 490 to insert the 10-mer. Also find 4 random positions to apply mutation to the 10-mer. Each 10-mer you have inserted should have 4 mutations.
- Your programs should take a file as input and a k value which will be the length of the consensus string. Run your algorithm for k=9, 10, 11 and comment about the strings in the project report. Both Randomized Motif Search and Gibbs Sampler are iterative algorithms. Let your Gibbs Sampler continue until the score of the algorithms no longer improve. For example, check your score every 50 iterations. If you see that the score remains the same for the last 50 iterations, then you can stop your algorithm.
- At the end of the programs, you will have 10 motifs. Use these motifs to find the consensus string. You can use the link below to draw the consensus strings. You need to give your motifs as the sequence data. http://weblogo.threeplusone.com/create.cgi
- You will prepare a project report that will include the project details and conclusions about the results.