# CS723/823 Introduction to Bioinformatics

# Project

# Fall 2017 – Assigned 11/2/17, Due 12/7/17

This project has two components. The first component is designed for you to explore a few current advances in understanding whole genomes. There are many entire genomes sequenced and the complete sequences of the genomes are archived in public databases. They are often annotated with many details regarding the knowledge at different locations of the genomes. The second component is designed for you to realize the in-depth need of computational algorithms in understanding genomic sequences.

**Task 1 – Accessing, extraction, and annotation of genomic sequences from chromosome maps. (20 points)**

There are two popular genome browsers . One is NCBI Genome Data Viewer at <https://www.ncbi.nlm.nih.gov/genome/gdv/>

Help page: <https://www.ncbi.nlm.nih.gov/genome/gdv/browser/help/>

The other is EpiGenome viewer at <http://epigenomegateway.wustl.edu/>

Question 1 – Briefly describe major functions provided by each of the two browsers. What is in common, and what is the difference between the two browsers?

Question 2 – Use NCBI Genome Viewer to perform the following function. You may need to read the help page provided regarding the use of the browser.

Select three consecutive genes on a chromosome. Each gene has a unique ID and a unique location to indicate the location of the gene on the chromosome. You may use any three consecutive genes from any chromosome of human. Report the following information.

(a). The name and location on the chromosome for each of the three genes.

(b). Download the segment of DNA sequence from the chromosome that contains the three genes and any sequence in-between two genes. You may use FASTA format or another format to represent the sequence.

(c). Indicate the location of the segment of chromosome sequence you downloaded in (b), so that it is easy to locate the segment in Genome Data Viewer. Describe briefly how this segment can be retrieved from the Viewer.

(d). In the file (FASTA) of the DNA sequence you downloaded in (b), mark the beginning and ending position on the sequence, so that it is clear which region of the sequence is a gene and which region is not.

**Task 2 - Implement the global alignment dynamic programming algorithm using an affine-gap penalty. (20 points)**

The input to your program involves two DNA sequences, a gap score for opening a new gap, and a gap score for extension of a gap, a score for a match and a mis-match respectively. The output includes the three matrices F(I,j), Ix(I,j), and Iy(I,j), the optimal alignment score, and the optimal alignment.

Report of the project

In the report, you are expected to submit your answers to Task 1 with any files needed. You are expected to submit a brief description regarding how to compile and run your program. If you want to provide an example, include the data in your submission. You are expected to submit the code for Task 2. You may use any programming language to complete the project. Submissions are on Blackboard.