

CSC 314, Final Project

Spring 2017

Bioinformatics is the study, development, and utilization of computational methods for storing, retrieving and analyzing biological data. The field of bioinformatics includes both the *development* of databases and tools for carrying out bioinformatics analyses and the *application* of these tools to answer important biological questions. I hope that you will leave this course with an appreciation of both the development and application of these tools.

For your Final Project, you will select an assignment related to either the development of a bioinformatics tool (i.e., a programming project), or the application of the databases and tools discussed during the semester to answer important biological questions. Your choice should be based on your interest and level of comfort with the project. You may work with a partner on the final project.

Final Project: select ONE assignment from either Option A or Option B.

Option A. *Bioinformatics programming project*

Write a bioinformatics program, in the language of your choice, that does one of the following. You must send me the source code for your program and I must be able to compile and run the program for you to receive credit. All source code used must be your own; library packages may be used only with permission.

1. A DNA or RNA sequence is read from a file. The program translates the entire sequence using all six possible reading frames. In addition, all open reading frames identified are highlighted. Note: you may not use Biopython or other available packages for this assignment.
2. Write a program that finds the optimal pairwise alignment between two sequences specified by the user. (You may assume that the sequences are no more than 100 characters each). Your program should output the score for the optimal alignment as well as the optimal alignment. The alignment is scored based on matches being worth 3 points, mismatches worth -1 points, and a linear gap penalty of 4. If multiple alignments are optimal, only one optimal alignment needs to be displayed. You may decide which alignment type (global, semi-global, or local) alignments you want to find.
3. Write a simple prokaryotic gene prediction program (pg 12 of the Gene Prediction notes), based on the promoter, Shine-Delgarno sequences, and open reading frames coding for at least 100 amino acids. Note: you may use Biopython for this assignment.
4. Using Biopython, write a script to BLAST the three sequences that we looked at from Lab 9, using the same parameters, and return the organism, accession number, similarity, e-value, and alignment of the top 3 hits.
5. You may choose to develop an additional bioinformatics program, with my approval.

Option B. *Bioinformatics analysis*

You will be given a gene to perform a bioinformatics analysis on. Your analysis may include but is not limited to the following tasks (the specific requirements will depend on the assigned gene).

- Draw a graph of the gene, labeling the positions of its introns and exons, and identifying the chromosome the gene is on.
- Identify similar proteins in other species, and whether the protein has any conserved domains.
- Based on the domain information, describe the function of the gene.
- Is the gene differentially expressed (I will tell you what experiments (i.e., GEO series) to look at)?
- Are mutations in the gene associated with any diseases or conditions?
- What secondary structure characteristics are present in the protein?

Important Dates

By 5:00 PM, April 26: On Piazza, post the following information:

1. If you will be working individually or with a partner, and if so, who
2. Your choice of Option A or Option B. If Option A, also state the programming project you will complete and the language you will write the program in.

May 12th, 4:00 PM: Final Projects are due through Blackboard (Option A or Option B) or hardcopy (option B only).

Additional Information

See the accompanying rubric for additional information.