Lab 3: Gene Density

Due date: October 14/ October 15.

About the Lab

This lab essentially repeats the software engineering and problem-solving process you've gone through for **Lab 2-2**, while introducing a few new wrinkles in it:

- The subject matter of the lab extends the subject matter of Lab 2. Your BIO 441 partners will get a research question for their quarter-long project and you will need to help them with the first stage of the project.
- You will start working with a new data file format, GTF[1] files. A GTF file contains the annotation of coding sequences in a given DNA fragment.

The lab is designed to span two lab periods: October 9 and October 14, with the final deliverables due October 15. The tentative outline of the software engineering process is as follows:

Date(s)	Stages
October 9 lab	Requirements and Design
October 9-15	Implementation (and redesign)
April 14 lab	Testing and delivery
April 15	Final delivery, submission

[1] Formerly known as GFF.

BIO 441 quarter-long project

The following is a general overview of the quarter-long project that is provided to your BIO 441 partners by Dr. Goodman (it is a verbatim quote from their assignment).

> *We are studying 4th chromosome (aka Dot chromosome and Muller element F) of Drosophila species. There are a few characteristics that make it interesting: it is very small in most fly species (D. ananassae is a notable exception, we may try to annotate some genes from it) and packaged primarily into heterochromatin, which tends to be transcriptionally silent (no or very little transcription). Yet this chromosome in D. melanogaster contains approximately 80 genes and many of them are expressed. In our study, we will compare Dot chromosome to a similar sized regions on another chromosome (3L) from the same species. We will try to characterize differences in sequence and gene content between two genome regions in the hope of finding what genome features may direct chromatin packaging (Dot into heterochromatic state) and allow gene expression from heterochromatin*[2].

Dr. Goodman provided BIO 441 with specific research questions and instructions on what needs to be computed in order to answer those questions. It is the responsibility of your BIO 441 partners to provide you more information about those needs.

Because the project sets forth a research question, different teams may formulate software requirements that are somewhat different, and therefore may develop software which will differ from each other. As long as (a) your final software deliverable satisfies the overall goals of the lab as set by Dr. Goodman to BIO 441 students, and (b) satisfies the specific requirements provided to you by your BIO 441 partners, the programs submitted for grading will be allowed to diverge from each other.

Lab Assignment

Your assignment is to create a program that analyzes a given DNA sequence and a description of locations of coding regions in it, and produces information describing the coding region. While your program may be asked to produce a variety of output, computation of a number of gene content measures is one of the core objectives of the assignment. In addition, some of your prior software may be reused/repurposed in this lab as well.

[2] The term chromatin refers to the 3-dimentional structure of DNA molecules (and RNA and protein molecules packed with them). A heterocromatic DNA molecule/fragment is more "tightly packed" then a eurochromatic. This affects numerous properties of the DNA, first and foremost, the gene expression (i.e., how often the gene is used to produce proteins).

The input to your program will be provided as two files: one file containing the DNA sequence and one file containing the coding regions/gene annotation. The first file is in the FASTA format. The second file is in the GTF format, which is the standard genome annotation file format. GTF files contain multiple entries, each entry occupying a single row. Each entry is a description of one coding region found in the DNA sequence. Your BIO 441 partners are responsible for providing an explanation of the exact format of the GFF files to you.

The specific requirements regarding the output of your program will be provided to you by your BIO 441 partners. When examining the requirements, please make certain you understand what output needs to be generated. Negotiate/clarify any ambiguous requirements. Look for omissions in the requirements.

Note: When planning your work, please take into account that the intended use of the Tuesday, October 14 lab period is to *test* your program both in terms of accuracy and satisfaction of requirements, but also in terms of usability. Please make sure that your team comes to class with software that your BIO 441 partners can help you test. *Do NOT* rely on the October 14 lab period for significant implementation activities - the implementation stage should effectively be completed by then.

Testing. While your BIO 441 partners have been asked from the very beginning of the course to produce test cases and test the work of your programs, starting this lab, we are going to emphasize this aspect of the software development process in the joint work. This means that your BIO 441 partners will be expected to contribute more and more to the testing of the software produced by your group as the course progresses. While part of their training on building use cases for the acceptance testing stage of the process will come from the yet-to-be-delivered lectures, it is expected that your collaborative work on these assignments will also contribute in a significant way. In particular, you are encouraged to involve your BIO 441 partners into testing activities you are engaging for unit and system testing purposes in order to illustrate to them how testing is usually done.

Submission Instructions

There are two CSC 448 deliverables that will be graded: the program that you develop and deliver to your partners and the program documentation/usage instructions. In order to properly evaluate your program though, we need you to submit two additional deliverables:

- The initial requirements document (as you see it at the beginning of the class);
- The final requirements document (as you modify it throughout the class);

(if your requirements document did not change, submit it twice and indicate that no changes were made).

Submission via handin. The full set of deliverables (source code, compilation/running instructions (README), user documentation, and the two requirements documents) needs to be submitted via the following `handin` command.

```
$handin dekhtyar 448-lab3 <files>
```

The submitted `README` file must identify all team members (and contain the team name if you have it). The deliverables submitted via `handin` are the ones that will be graded.

Submission via Piazza. You must deliver the final (working!) version of your software to your BIO 441 partners via Piazza. While your BIO 441 partners are responsible for posting the two (initial, final) versions of your team's requirements documents there (as well as any test cases/input data on which the program needs to run), the CS side of the team is responsible for posting the final executable/runnable program, and user instructions.

Deadlines. It is expected that each team will perform initial software delivery to the BIO 441 students by the end of the October 14 lab. Please make sure the initial Piazza deliverables are up by then.

However, we are cognizant of the fact that your BIO 441 partners may not have enough time to work with your final version of the software. Because of this, there is a 24-hour *grace period* in effect. During this grace period, you may work on any bugs/issues noticed by the BIO 441 partners and replace the old deliverables on Piazza (and the matching source code you submit via `handin`) with the new ones.

The official **hard deadline** to have the ultimate Piazza and `handin` submissions is *October 15, 11:59pm*.

Good Luck!