Syllabus
Course Title

Algorithms for DNA Sequencing

Course Instructor(s)
Ben Langmead and Jacob Pritt

Course Description
We will learn the concepts, algorithms and data structures that are foundational to the analysis of DNA sequencing data. We will use what we learn, together with our programming skills, to implement these and similar ideas. We will work with real data: real genome sequences and real sequencing data.
To succeed in this course, it is helpful to have some undergraduate-level background in computer science, though we have worked hard to make the course accessible even to students with minimal or no computer science background. You definitely *do not* have to be a trained biologist. You don't have to be an expert on genomics and genomes. You don't have to know how a DNA sequencer works. You just have to be curious about these things, and to be open to learning a little about genomics and sequencing as you go.
You will need Python programming skills. All of the code that we'll show you in class, in the lecture notes and in the practical sessions, will be written in Python. We will sometimes ask you to modify or adapt our examples written in Python. So if you're not so confident in your Python skills, you might want to get some more practice before you enroll, or you might go ahead and enroll, but if you do you should set aside some extra time to learn and practice as you go along.

Course Content
Module 1: DNA sequencing, strings and matching: Why study this? DNA sequencers and how they work. How DNA can be represented as a string. Using Python to parse and manipulate real genome sequences and real DNA sequencing data. Naive exact matching.
Module 2: Preprocessing, indexing and approximate matching: Improving on naive exact matching with Boyer-Moore. Preprocessing and indexing. Indexing: grouping and ordering. *k*-mers and *k*-mer indexes. Approximate matching and the pigeonhole principle.
Module 3: Edit distance, assembly, overlaps: Hamming and edit distance. Algorithms for computing edit distance. Dynamic programming. Global and local alignment. De novo assembly. Overlaps and overlap graphs.
Module 4: Algorithms for assembly: Shortest common superstring and the greedy version. How repetitive DNA makes assembly difficult. De Bruijn graphs and Eulerian walks. How real assemblers work. The future of assembly. Wrap up.
Weekly quizzes
There are four weekly quizzes. You may begin submitting them as soon as the course opens. Quiz 1 is due at the end of the first week, Quiz 2 is due at the end of the second week, Quiz 3 is due at the end of the third week, and Quiz 4 is due at the end of the fourth week.
Quiz Scoring
You may attempt each quiz up to 3 times in 8 hours. The score from your most successful attempt will count toward your grade.
Grading policy
You must receive a final grade of 70% or better on each assignment (quizzes and homework) to pass the course.

Your final grade will be calculated as follows:

Quiz 1 = 10%

Quiz 2 = 10%

Quiz 3 = 10%

Quiz 4 = 10%

Programming Homework 1 = 15%

Programming Homework 2 = 15%

Programming Homework 3 = 15%

Programming Homework 4 = 15%

Differences of opinion
Keep in mind that currently data analysis is as much art as it is science - so we may have a difference of opinion - and that is ok! Please refrain from angry, sarcastic, or abusive comments on the message boards. Our goal is to create a supportive community that helps the learning of all students, from the most advanced to those who are just seeing this material for the first time.
Plagiarism

Johns Hopkins University defines plagiarism as "...taking for one's own use the words, ideas, concepts or data of another without proper attribution. Plagiarism includes both direct use or paraphrasing of the words, thoughts, or concepts of another without proper attribution." We take plagiarism very seriously, as does Johns Hopkins University.

We recognize that many students may not have a clear understanding of what plagiarism is or why it is wrong. Please see the following guide for more information on plagiarism:

http://www.jhsph.edu/academics/degree-programs/master-of-public-health/current-students/JHSPH-ReferencingHandbook.pdf

It is critically important that you give people/sources credit when you use their words or ideas. If you do not give proper credit -- particularly when quoting directly from a source -- you violate the trust of your fellow students. The Coursera Honor code includes an explicit statement about plagiarism:

I will register for only one account. My answers to homework, quizzes and exams will be my own work (except for assignments that explicitly permit collaboration). I will not make solutions to homework, quizzes or exams available to anyone else. This includes both solutions written by me, as well as any official solutions provided by the course staff. I will not engage in any other activities that will dishonestly improve my results or dishonestly improve/hurt the results of others.